1. Modeling Behavior Data in Security and Defense Applications

Robert G Abbott, Sandia National Labs

Realistic models of human behavior have many applications in national security and defense, such as training fighter pilot tactics, force-on-force physical security simulations, and modeling cyber threats. Despite the seeming dissimilarity between these applications, each requires simulated role players to stand in for adversaries, allies, and neutral bystanders. Developing a behavior model is labor-intensive, which is especially problematic in fast-evolving competitive domains. Using data-driven approaches, we can analyze and simulate behavior in each case. The general steps are the same — instrument the environment, collect examples of the desired behavior, label/score the data, and use machine learning to create a computational model. While the proliferation of low-cost sensors might imply this is a Big Data problem, in our experience many defense-oriented applications present significant obstacles to data collection. These include a small, tightly knit operational community and stringent security considerations. As a result, data is often scarce, and our models must incorporate heuristics and subject matter expertise along with the behavior traces.

Abbott, R.G.; Lakkaraju, K.; Warrender, C. Transitioning from Human to Agent-Based Role-Players for Simulation-Based Training. HCII 2015. Aug 2-7 2015. Los Angeles, CA.

2. Adaptive Impact-Driven Detection of Silent Data Corruption for HPC Applications

Sheng Di, Argonne National Laboratory

For exascale HPC applications, silent data corruption (SDC) is one of the most dangerous problems because there is no indication that there are errors during the execution. We propose an adaptive impact-driven method that can detect SDCs dynamically. The key contributions are threefold. (1) We carefully characterize 18 real-world HPC applications and discuss the runtime data features, as well as the impact of the SDCs on their execution results. (2) We propose an impact-driven detection model that does not blindly improve the prediction accuracy, but instead detects only influential SDCs to guarantee user-acceptable execution results. (3) Our solution can adapt to dynamic prediction errors based on local runtime data and can automatically tune detection ranges for guaranteeing low false alarms. Experiments show that our detector can detect 80-99.99% of SDCs with a false alarm rate less that 1% of iterations for most cases. The memory cost and detection overhead are reduced to 15% and 6.3%, respectively, for a large majority of applications.

Joint work with Franck Cappello (Argonne National Laboratory)

3. Framework for Error Modeling in ALE Numerical Trials (FERMANT)

Yuriy Ayzman, Texas A&M University, Lawrence Livermore National Laboratory

Multiphysics codes rely on Arbitrary Lagrangian Eulerian (ALE) relaxers to resolve mesh tangling. These relaxers introduce numerical error due to mixing across boundaries. Current methodology for determining simulation quality is whether the run completes successfully. In order to better quantify the introduction of ALE error, problem specific quantities were tied to ALE driven ones. The result is a framework for judging the impact of ALE on simulation accuracy. This framework helps quantify the amount of degradation in computation results.

Joint work with Dan Laney, Josh Kallman

4. Dakota: Algorithms for Design Exploration and Simulation Credibility

Brian M. Adams, Sandia National Laboratories

Sandia National Laboratories' Dakota software (http://dakota.sandia.gov) delivers advanced parametric analysis techniques enabling quantification of margins and uncertainty, risk analysis, model calibration, and design exploration with computational models. Its methods include optimization, uncertainty quantification, parameter estimation, and sensitivity analysis, which may be used individually or as components within advanced strategies to address questions like 'What is the best design?', 'How safe is it?', and 'How much confidence do I have in my answer?' Available publicly under an open source license, Dakota is used broadly by academic, government, and corporate institutions and plays a critical role in DOE ASC, Office of Science, and Office of Nuclear Energy Programs. This poster will survey Dakota capabilities, emphasizing data-driven approaches such as deterministic and Bayesian calibration, surrogate (response surface) modeling, and sequential design of computational experiments.

Joint work with Adam Stephens, Sandia National Laboratories, Patricia Hough, Sandia National Laboratories

---

5. Multiscale Imaging of Carbonate Rocks and Efficient 3-D Stochastic Rock Reconstruction through Dimension Reduction

Jonghyun Harry Lee, Stanford

Geomaterials containing nano-pores (e.g., shales and carbonate rocks) have become increasingly important for emerging problems such as unconventional gas and oil resources, enhanced oil recovery, and geologic storage of CO2. Accurate prediction of coupled geophysical and chemical processes at the pore scale requires realistic representation of pore structure and topology. This is especially true for chalk materials, where pore networks are small and complex, that need to be characterized at sub-micron scale.

We apply integrated multiscale imaging of carbonate rock to characterize 3-D structures and mineral distributions for petrophysical and mechanical properties from nanometer to centimeter scales. In particular, laser scanning confocal microscopy is used to characterize pore structures and microlithofacies at a micron resolution over the sample of thin-section (~ 2-3 cm) scale and dual focused ion beam-scanning electron microscopy (FIB-SEM) to image 3D nanometer-to-micron-scale microcracks and pore distributions. With a limited sample volume (~10's micron) from imaging techniques for nano-pore characterization, it is critical to define representative sampling of FIB-SEM images and apply it to the thin-section or larger scale. In this work, several texture characterization techniques are applied for segmentation clusters represented by a couple of 3D FIB-SEM image stacks per each cluster. Representative 3D pore structure at a FIB-SEM scale is characterized using multi-point statistics and the ensemble member of 3D stochastic pore structures are generated. These ensemble members are then projected onto the primary feature-based subspace via principal component analysis in order to reduce the size of the ensemble. The reduced dimension of stochastic ensemble members will be used to reconstruct digital rock pore networks at the centimeter scale to represent heterogeneity of nano-porous geomaterials. In addition, these multiscale sets of reconstructed digital rocks will be used to obtain permeability at several different scales. Upscaling of permeability to the Darcy scale (e.g., the thin-section scale) with image datasets will be discussed with emphasis on understanding microfracture-matrix interaction, representative volume for FIB-SEM sampling, and multiphase flow and reactive transport.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Joint work with Hongkyu Yoon, Geomechanics Department, Sandia National Laboratories, Albuquerque, NM, 87185 USA, Peter K. Kitanidis, Civil and Environmental Engineering, Stanford University, Stanford, CA, 94305 USA

---

6. Scalable Subsurface Inverse Modeling Using Big Environmental Data Sets: Application to Conductivity Estimation in a Sand Box from a Massive MRI Data Set

Jonghyun Harry Lee, Stanford

Characterizing subsurface properties, particularly hydraulic conductivity, is crucial for reliable and cost-effective groundwater supply management, contaminant remediation, and emerging deep subsurface activities such as geologic carbon storage, unconventional resources recovery, and induced seismicity. With recent advances in sensor technology, a large volume of hydro-geophysical and chemical data can be obtained to achieve high-resolution images of subsurface properties, which can be used for accurate subsurface flow and reactive transport predictions. However, subsurface characterization with a plethora of information requires high, often prohibitive, computational costs associated with big data processing and numerous large-scale numerical simulations. As a result, traditional inversion techniques are not well-suited for problems that require coupled multi-physics simulation models with big data sets.

In this work, we apply a scalable inversion method called Principal Component Geostatistical Approach (PCGA) for characterizing heterogeneous hydraulic conductivity (K) distribution in a 3-D sand box. The PCGA is a Jacobian-free geostatistical inversion approach that uses the leading principal components of the prior covariance to reduce computational costs, sometimes dramatically, and handle a large number of measurements (e.g. 10^6 or more) by constructing a fast preconditioner whose computational cost scales linearly with the data size. The optimal number of principal components to be used in the inversion is determined through generalize Hermitian eigenvalue analysis. Sequential images of transient tracer concentrations in the sand box were obtained using magnetic resonance imaging (MRI) technique, resulting in 6 million tracer-concentration data. Since each

individual tracer observation has little information on the K distribution, the dimension of the data was reduced using temporal moments. Consequently, 100,000 unknown parameters (i.e., spatial K distribution) consistent with the scale of MRI data (at a scale of 0.25^3 cm^3) were directly estimated by matching temporal moments the original tracer data. Only about 2,000 forward simulations in total were required to obtain the best estimate with corresponding estimation uncertainty, and the estimated K field captured key patterns of the original packing design, showing the efficiency and effectiveness of the proposed method. This demonstration shows that PCGA can be a practical tool for characterizing multiphysics processes involving expensive forward simulations and a large amount of data sets.

Joint work with Hongkyu Yoon, Geomechanics Department, Sandia National Laboratories, Albuquerque, NM, 87185 USA, Peter K. Kitanidis, Civil and Environmental Engineering, Stanford University, Stanford, CA, 94305 USA

---

7. Information-Preserving Data Reduction for Scientific Applications through Efficient Data Modeling
Yu Zhuang and Misha Ahmadian, Texas Tech University

In many science and engineering investigations, there are data that are of critical importance but highly expensive to generate, e.g. obtainable through time-consuming experimental or computational processes. There is a strong incentive to reduce the amount of such data, since reduction of expensively generated data also means reduction of data generation costs. But what is critical is that the data reduction should not lead to, within a problem-dependent threshold, the loss of information carried by the original larger data set. Thus, for scientific studies with large expensive datasets, this poster presents an inexpensive data modeling procedure to replace the generation of some, hopefully as much as possible, data. The data modeling procedure described here is made to be accurate using machine learning techniques so that the modeled data carries almost the same scientific information as the original expensive data. The data modeling procedure uses data generated earlier in the investigation process to approximate data needed at the current moment in the investigation process, leading to the reduction of expensive data generation operations through increased data utilization. The data modeling procedure was applied to ab initio chemical dynamics simulations by storing, managing, and utilizing the highly expensive ab initio potential energy data, in order to model potential energy information needed in additional steps of the simulation. This modeling procedure shows increased efficiency of the dynamics simulation, while maintaining the simulation accuracy.

Joint work with William L. Hase (Dept of Chemistry & Biochemistry,Texas Tech University), Yong Chen (Dept of Computer Science, Texas Tech University)

---

8. Spatial Context Driven Manifold Learning for Image Feature Extraction
Dalton Lunga, Oak Ridge National Laboratory, GIST Group

Manifold learning methods have demonstrated with confidence, the representation of spectral signatures for hyperspectral dimensional images. This has proven to be beneficial when working with images that consist of spectral features with very subtle differences and often spatially induced disjoint classes. For such images, their neighborhood relations are difficult to capture using traditional graph based embedding techniques because robust parameter estimation is a challenge for custom kernel functions that compute neighborhood graphs e.g finding the optimal number of nearest neighbors. Our ongoing efforts have been to address these challenges by studying spatial context driven manifold learning methods. Building upon empirical findings, the study has been able to reveal that the use of spatial contextual information has a bearing on the structure of the graph Laplacian which in turn links image pixel observations to their manifold spaces. Further experimental results demonstrate an improvement in the classification performance compared to traditional kernel based manifold learning methods. As such, our overall long term efforts are towards scaling context driven manifold learning methods towards feature extraction for very large images.

Joint work with Lexie Yang, Oak Ridge National Laboratory, GIST Group; Saurabh Prasad, Department of Electrical and Computer Engineering, University of Houston; Melba Crawford, School of Civil Engineering, Purdue University

---

9. Statistical Distributions of Spectral Anomalies in Varied Scene Content
Emily E. Berkson, Rochester Institute of Technology

Anomaly detection (AD) algorithms are frequently applied to hyperspectral imagery, but different algorithms produce different outlier results depending on the image scene content and the assumed background model. Outliers in statistical AD algorithms should theoretically approximate a chi-squared distribution; however, this is

rarely the case with real hyperspectral imagery. The expected distribution of outliers found with graphical methods remains unclear. In this work we quantify differences in the distributions of anomalous pixels found with two multivariate statistics-based algorithms (RX and subspace-RX), and one graphical algorithm (TAD). We also look for general trends in performance with varied scene content, where we expect cluttered scenes to have more outliers of a different statistical distribution than simple scenes. Three separate scenes were extracted from the hyperspectral MegaScene image taken over downtown Rochester, NY with the VIS-NIR-SWIR ProSpecTIR instrument. In order of most to least cluttered, we study an urban, suburban, and rural scene. Each scene was run through the three aforementioned AD algorithms, and the distributions of the most anomalous 5% of pixels were compared. Results show statistical characterizations of anomalies detected by the three methods, applied to varying scene content. This work will lead to a deeper understanding of these algorithms and their applicability across different types of imagery.

Joint work with David W. Messinger (Ph.D.)

---

10. Multivariate Curve Resolution for Hyperspectral Images

Stephen M. Anthony, Sandia National Laboratories

Hyperspectral imaging extends the spatial resolution of traditional imaging to the spectral domain, measuring the entire emission spectrum for each spatial position. The resulting hyperspectral data cubes contain a wealth of data, theoretically allowing the determination of both the underlying spectral components and their spatial distribution. As a result, hyperspectral imaging has numerous applications in such diverse fields as biomedical imaging and satellite remote sensing. However, in practice determining the spectra and concentrations can be quite challenging. Often only limited prior information is available and samples can include spectrally similar compounds. Three different classes of algorithms, principal component analysis (PCA), independent component analysis (ICA), and multivariate curve resolution (MCR), all offer potential solutions. While each algorithm will decompose the hyperspectral data cube into spatial and spectral components, only MCR can ensure physically reasonable results (such as forbidding negative concentrations). Unfortunately, MCR suffers from a number of limitations, most notably that there may not be a single unique solution due to rotational ambiguity. Further, most MCR implementations assume independent and identically distributed (iid) noise. This assumption is invariably violated when acquiring hyperspectral Raman or fluorescence images of living cells, where the Poisson shot noise causes a signal-dependent noise distribution. Here, I am developing an MCR implementation addressing these limitations. In addition to accounting for the noise distribution, my implementation is designed to minimize the uncertainty in the final solution where possible. When eliminating the rotational ambiguity is impossible, my implementation will quantify the uncertainty, providing feasibility bands for the spectra.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

---

11. Joint Geophysical and Flow Inversion to Characterize Fracture Networks in Subsurface Systems

Maruti Kumar Mudunuru, Los Alamos National Laboratory

Subsurface applications including geothermal, geological carbon sequestration, oil and gas, etc., typically involve maximizing the extraction of energy or maximizing the storage of fluids. Characterizing the subsurface is extremely complex due to heterogeneity, and anisotropy. There are uncertainties in the geological, geomechanical, geochemical, thermal, hydraulic parameters, which need to be estimated from multiple diverse as well as fragmented data streams. State-of-art inversion approaches do not consider multiple data sets as well as confidence in experimental/field-scale measurements and models used. Joint inversion methods attempt to overcome these limitations by combining different sources of information and inverting multiple types of data residuals simultaneously. We present a coupled multi-physics framework, for integrating data from geophysical and flow sources, and constraining fractured subsurface system characteristics such as permeability. The subsurface system is modeled as a Discrete Fracture Network (DFN) where fractures are assumed to be two-dimensional planes in three-dimensional space. Using a two-step joint inversion approach, we then evaluate and constrain the stochastics of the DFN based on microseismic and flow datasets. In this approach, we first estimate bounds on the statistics for the DFN fracture orientations and fracture sizes, using microseismic data. Then, using a Monte Carlo sampling based on these estimated fracture statistics, multiple realizations of DFNs are generated. On these realizations, we perform flow calculations, and compare with measured flow datasets, to obtain optimal probability distribution of the DFNs. The obtained DFNs can then be used to evaluate subsurface parameters of

interest such as permeability and porosity, which can be eventually used for predictive modeling of state-of-stress due to anthropogenic activities, for instance. The efficacy of this multi-physics based joint inversion is demonstrated with representative synthetic examples.

Joint work with Satish Karra, Los Alamos National Laboratory

---

12. Application of Metaheuristic Optimization Methods for Neutron Spectral Shaping Applications
James Bevins, University of California, Berkeley (UCB)

Neutron sources are broadly classified by their intensity, or flux, and energy distribution. For many applications, such as medical treatments, radiation damage studies in fast reactors, nuclear effects, or nuclear forensics, there are no operational neutron sources that have the desired intensity and energy distribution. Historically, surrogate methods, equivalencies, and alternative sources were developed to provide calibration metrics that allowed the use of different neutron and ion sources to meet research and use objectives in these areas. To provide more effective neutron sources, this research develops energy tuning assemblies (ETAs) to tailor the neutron spectra of existing sources to produce desired neutron spectra where capability gaps exist.

Successful manipulation of neutrons is achieved through selectively choosing materials with the right nuclear physics properties, which are encapsulated in nuclear data libraries. However, there are 83 elements containing 292 stable or long-lived isotopes, each with up to dozens of nuclear reactions that can affect the neutron spectra, available for use in spectral shaping. Managing, searching, and manipulating this nuclear data is crucial to accomplishing any neutron spectral shaping goals.

Because of the volume of materials and interaction mechanisms, most spectral shaping to date has relied on one reaction of one isotope. Examples include neutron filters such as Cadmium, low-Z elastic scattering to soften spectra, nuclear reactions to create mono- or nearly-mono-energetic spectra, or resonances to create transmission spectra. More recently, research done in support of Boron Neutron Capture Therapy (BNCT) started to scratch the surface of what is possible when nuclear data is fully leveraged to shape a spectrum.

However, most of the BNCT work was done using trial and error with a small selection of materials [1]. Optimization attempts were made, but those efforts were incremental improvements to existing designs rather than searches for revolutionary design improvements [2]. In contrast, this research develops META-CODE (METAheuristic based Cuckoo search Optimization algorithm for Designing Etas) to effectively search the nuclear data and design space. META-CODE uses PARTISN to perform the radiation transport calculations to generate output spectra [3]. A design fitness is assessed by comparison of the output and objective spectra using U-optimality criteria. Cuckoo Search (CS)-based algorithms are then used to search for designs with greater fitness [4].

CS algorithms combine many of the best metaheuristics from other swarm intelligence and evolutionary algorithms. Like the use of these search and optimization techniques in other fields, META-CODE will allow for new capabilities to be extracted from old data. This will result in novel neutron spectral shaping approaches that can be designed for many divergent applications and constraints more effectively than ever before.

Joint work with Rachel Slaybaugh, UCB, Lee Bernstein, LLNL, LBNL, UCB, Bill Dunlop, LLNL, Gene Henry, LLNL

[1] M. Asnal, T. Liamsuwan, et al., An Evaluation on the Design of Beam Shaping Assembly Based on the D-T reaction for BNCT, J. Phys. Conf. Ser., vol. 611, 2015.
[2] J. Vujic, E. .Greenspan, et al., Optimal Neutron Source & Beam Shaping Assembly for Boron Neutron Capture Therapy, UCBNE-4239, 2003.
[3] R. E. Alcouffe, R. S. Baker, et al., PARTISN: A Time-Dependent, Parallel Neutral Particle Transport Code System, LA-UR-08-07258, 2008.
[4] X.-S. Yang and S. Deb, Cuckoo Search via Levy Flights, in Nature & Biologically Inspired Computing, 2009, pp. 210-214.

---

13. A Machine Learning Approach for Automatic Characterization of Memory Faults
Lissa Baseman, Los Alamos National Laboratory

As high-performance computing systems continue to grow in scale and complexity, the study of faults and errors is critical to the design of future systems and mitigation schemes. One key aspect of memory reliability frequently investigated is fault mode type in system DRAM. While current schemes require offline analysis for proper classification, current state-of-the-art mitigation techniques require accurate online prediction for optimal performance. In this work, we explore the prediction performance of an online machine learning-based approach in classifying memory fault modes from two leadership-class supercomputing facilities. Our results compare the

predictive performance of this online approach with the current offline approach. We also report on those features found most important to predictive performance. Finally, we investigate the universality of our classifiers by evaluating prediction performance using training data from disparate computing systems. This work provides a critical analysis of this online learning technique and can benefit system designers to help inform best practices for dealing with reliability on future systems.

Joint work with Nathan Debardeleben (Los Alamos National Laboratory), Kurt Ferreira (Sandia National Laboratory), Scott Levy (Sandia National Laboratory), Steven Raasch (AMD), Vilas Sridharan (AMD), Taniya Siddiqua (AMD), Qiang Guan (Los Alamos National Laboratory)

---

14. How an Ensemble of Simulations Informs Our Understanding of NIF Implosions

Ryan Nora, Lawrence Livermore National Laboratory

The achievement of inertial confinement fusion ignition on the National Ignition Facility relies on the collection and interpretation of a limited (and expensive) set of experimental data. This inherently uncertain data is obtained from a variety of diagnostics upon which we use to infer and predict the behavior of the imploding capsule. This data is therefore supplemented with state-of-the-art multi-dimensional radiation-hydrodynamic simulations to provide a better understanding of how the implosion proceeds. We present a relatively large number (~1500) of systematically perturbed two-dimensional simulations and their corresponding synthetic experimental images to verify our understanding of low-mode fuel and ablator asymmetries seeded by asymmetric illumination. The ensemble of simulations fills a seven-dimensional space, which is decomposed with a polynomial chaos expansion. This provides a quantitative comparison with experimental data as well as shedding insight to physical questions such as scaling-surrogacy issues (between sub-scale and full-scale implosions) and capsule robustness (regarding asymmetric stagnation mitigation strategies).

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-ABS-681578

J. Luc Peterson, J. E. Field, B. Spears, S. Langer, S. Brandon, A. Pak, P. T. Springer, J. Gaffney, J. H. Hammer, M. Buchoff, A. Kritcher

---

15. Project Morpheus: Exploiting Next-Generation Supercomputer Architecture and Scalable In-Transit Data Analysis to Design a Robust Inertial Confinement Fusion Implosion

J. Luc Peterson, Lawrence Livermore National Laboratory

Inertial confinement fusion experiments, such as those fielded at the National Ignition Facility at Lawrence Livermore National Laboratory, aim to achieve fusion energy gain by the spherical implosion of thermonuclear fuel. Since the requisite > 30x compression precludes perfectly spherical implosions, there exists great interest in experimental designs that can produce significant thermonuclear yield even in the presence of realistic asymmetrical perturbations. However, the desire to use computer simulations to search for and design a robust implosion is weighed by the difficulty in doing so: the design space is large and each simulation is costly. A meaningful search likely requires at least 10-20 dimensions, with each run taking hundreds of CPU-hours to produce roughly 1 TB of raw data. Furthermore, the structure of the search space is unknown a priori, so the number of samples needed to adequately build models is uncertain.

To tackle this problem, we are developing in-transit data analysis that exploits the availability of fast, local memory on next-generation supercomputers, such as the burst-buffers on the new Trinity machine at Los Alamos National Laboratory. As a simulation writes its raw data to this fast memory, it sends a signal to a queue server that data are available for independently running analysis nodes to examine. The analysis nodes find the raw data, distill it to physically interesting quantities (such as integrated metrics of theoretical importance or synthetic data that can be directly compared to experiments), compress and transfer the results to disk, and erase the raw data before the fast memory overflows. Uncertainty Quantification software directs the execution of individual simulations and builds Polynomial Chaos Expansions of the integrated metrics, while a stratified hypercube sampling scheme allows new dimensions to be explored as sub-spaces converge. During the Trinity Open Science campaign (beginning Feb 2016), we expect to run roughly 200k individual simulations and generate 50 TB of compressed data. Hopefully the Morpheus Data Set will shed light on not only how asymmetries affect implosions but also how to make them more resilient.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Joint work with S. Brandon, J. Gaffney, J. E. Field, S. Langer, R. Nora, B. K. Spears (Lawrence Livermore National Laboratory)

---

16. Simulation and Analysis of Inertial Confinement Fusion Experiments
Jim A Gaffney, Lawrence Livermore National Laboratory

Inertial confinement fusion experiments aim to compress and heat hydrogen to conditions at which nuclear fusion reactions ignite and release energy. If successful, fusion ignition will pave the way to near limitless, inherently safe, and clean energy. Current experiments, in which millimeter-scale spherical capsules filled with hydrogen isotopes are compressed by 100x under the action of some of the world's largest laser systems, present a severe data analysis challenge for multiple reasons. Experimental data are exceptionally expensive, and the nature of the platform means that only indirect, integrated observations are possible; the space of controlling parameters has hundreds of dimensions; and the response of the capsule is highly nonlinear. Alongside this, computer models of capsule evolution rely on multiple physics models, some of which are pushed into previously unexplored regimes of their respective parameter spaces. In this poster we will describe recent work at Lawrence Livermore National Laboratory to understand experimental results, based on a hierarchy of simulation methods and data-analytic tools.

Ongoing experiments at the National Ignition Facility aim to combine a suite of simulation methods varying from rapid approximate methods to large-scale multiphysics calculations with sparse experimental data in order to move current designs towards fusion ignition. We will discuss the challenges to this work, and present results from fast models in which energy transport and hydrodynamic flow are solved analytically. We aim to assess the information content, shot to shot consistency, and the utility in understand underlying physics of the available experimental data, and to make steps towards a consistent analysis that includes all of the available diagnostic quantities. We focus on several aspects of the problem, namely the poorly understood failure modes of current designs, possible problems with underlying physics models, and the highly integrated nature of observations.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-ABS-681699

Joint work with S. Brandon (LLNL), D. Callahan (LLNL), M. Buchoff (LLNL), J. E. Field (LLNL), J. H. Hammer (LLNL), T. Ma (LLNL), R. Nora (LLNL), J. Luc Peterson (LLNL), B. Spears (LLNL), P. T. Springer (LLNL)

---

17. A Bayesian Method of Determining Absolute Efficiency in Coincident Particle Detectors
David K. Olson, Los Alamos National Laboratory

Space plasma and energetic neutral atom instruments used to measure particle fluxes have a drawback in not being able to obtain an absolute calibration. Assumptions and simplified models are used instead to derive instrument calibrations. Instruments using coincident detection provide additional information that can be used to derive a reliable estimate of uncertainty on calibration, improving the determination of in-situ particle fluxes in orbit. We use a Bayesian model to analyze laboratory calibration data for the Helium, Oxygen, Proton, Electron (HOPE) mass spectrometer on NASA's Van Allen Probes mission to explain calibration data and derive an error estimate for this instrument. The model will be applicable to other coincident particle instruments and will lead to a more complete and accurate understanding of uncertainty in particle flux measurements.

Joint work with Brian A. Larsen, Brian P. Weaver

---

18. APT Malware Static Trace Analysis Through Bigrams and Graph Edit Distance
Alexander D. Bolton, Imperial College London

Malicious software, or malware, is a growing problem for the protection of networks and important information. Gaining understanding about malware samples is crucial in order to ensure cyber security at institutions such as Los Alamos National Laboratory (LANL). Advanced Persistent Threat (APT) malware, which is designed with a specific target in mind, is often generated by modifying existing malware. Malware samples that are versions of an original malicious program form a malware family. Reverse engineers, who seek to detect the origin and purpose of malware, can work more efficiently if they know the family that generated a piece of malware.

This paper proposes a two-part method for comparing a new sample to a database of existing malware samples categorized by family label. The first part is a fast filtering method to disregard samples that are dissimilar to the new sample, based on the instruction sequences generated by the samples. Once the filter has been applied, the static trace for each database sample and the new sample are converted to graphs with nodes representing subroutines and edges representing calls from one subroutine to another. The shortlisted static trace graphs are compared to the new sample using an approximation of the graph edit distance. A classifier based on the combination of the two methods can accurately place new malware samples into their appropriate families.

Joint work with Christine Anderson-Cook (Los Alamos National Laboratory)

---

19. Analyzing Multisource and Multidimensional Data to Evaluate Spatio-Temporal Trends and Patterns of Potential Impacts Related to Modelled Offshore Hazards
Jennifer Bauer, NETL/AECOM

Efforts to prepare for and reduce the risk of hazards, from both natural and anthropogenic sources, that threaten our oceans and coasts requires an understanding of the dynamics and interactions between the physical, ecological, and socio-economic systems. Understanding these coupled dynamics are essential to ensuring energy security as offshore oil & gas exploration and production continues to push into harsher, more extreme environments where risks and uncertainty both increase. However, evaluating the range of hazards posed to and by offshore energy exploration and production requires robust analyses capable of incorporating multi-source, multi-format, multi-scale, and multi-dimensional data. Here, we present spatio-temporal analytics methods for evaluating relationships between a broad range of multi-source, format, scale and dimensional anthropogenic and natural datasets to evaluate potential environmental, social, and economic impacts from modeled offshore hazard scenarios, such as blowouts or hurricanes. These methods stress approaches for combining spatio-temporal, multisource data, highlighting their ability to contribute to the rapid analysis and effective communication of analytical results to aid a range of decision-making needs, highlight key knowledge and/or technology gaps, and reduce uncertainty.

Joint work with Kelly Rose – NETL, Lucy Romeo - NETL/AECOM, Jake Nelson - NETL/ORISE, Dorothy Dick - NETL/ORISE

---

20. Probabilistic Forecasting of Seasonal Influenza in the U.S.
Dave Osthus, Los Alamos National Laboratory

Seasonal influenza is a serious public health and societal problem due to its consequences resulting from absenteeism, hospitalizations, and deaths. The overall burden of influenza is captured by the Centers for Disease Control and Prevention's (CDC) influenza-like illness (ILI) network, which provides invaluable information about the current incidence. Despite the relatively rich surveillance data and the recurrent nature of seasonal influenza, forecasting the timing and intensity of seasonal influenza in the US represents a significant challenge. This is because the form of the disease transmission process is uncertain, the disease dynamics are only partially observed, and the public health observations are noisy.

In this work, we model the data, disease transmission process, discrepancy, and parameters jointly in a Bayesian probabilistic model. This is a promising approach for forecasting seasonal influenza, as it allows us to impose constraints based on historical flu seasons. Accounting for multiple modes of uncertainty and systematic model discrepancy results in a probabilistic, prospective forecast of the current flu season, available at: bsvgateway.org/flu/

Joint work with Jim Gattiker (LANL), Reid Priedhorsky (LANL), Sara Del Valle (LANL)

---

21. Power System Dynamic Estimation
Thomas Catanach, California Institute of Technology

Because power systems are becoming increasingly complex and subject to disturbances, developing methods for state estimation and system identification is essential for increasing the reliability of the power grid. Currently this problem is solved on slower, steady state, time scales; however, with the deployment of phasor measurement units (PMUs) faster estimation is now possible. Many methods have been studied for dynamic state estimation including both local and global filtering methods. One of the main challenges for global methods is how the filter integrates the differential algebraic equations (DAE) that describe the power system. This work applies implicit methods used to solve DAEs to improve the performance and robustness of an Extended Kalman Filter, making it

an attractive state estimation choice. Further we introduce techniques to reduce the effect of temporary disturbances on the state estimated to help track the state through these faults where the network model is no longer accurate by creating a layered estimation architecture. This architecture integrates state estimation, change point detection, and classification of disturbances.

Joint work with Manuel Garcia, University of Texas, Scott Vander Wiel, LANL, Russell Bent, LANL, and Earl Lawrence, LANL

---

22. Calibration with Applications to Carbon Capture for Multiscale Models and Large Parameters
Sham Bhat, Los Alamos National Laboratory

Uncertainty quantification (UQ), and especially Bayesian calibration, help obtain reliable predictions for multi-scale problems. These models often must be constrained to experimental bench-scale data. The calibration framework requires discrepancy term between the model and reality, and usually a stochastic emulator trained to limited output from a computationally expensive model. Here we present advancements in statistical methodology in UQ and calibration with applications to many scientific problems. These include calibration of complex models to bench-scale data, accounting for due functional inputs and outputs, an intrusive dynamic discrepancy approach for upscaling of uncertainty, calibration problems with a large number of potential parameters, and an input space of potential functions. We used many flexible and computationally approaches for the emulator and/or discrepancy, including a Bayesian Smoothing Spline (BSS) ANOVA Gaussian Process. The methodology presented may have far-reaching impact in many areas of science where multiscale modeling is used.

---

26. Chemical Forensics of Uranium Oxide Processing: Using Chemical and Metallic Impurities to Predict Sample Provenance
Karl Pazdernik, North Carolina State University

Determining the origin of uranium materials is of great interest in forensic science. Here we seek to link the individual uranium oxide processing steps (precipitation temperature, calcination temperature, pH of precipitate, urea concentration, precipitate method, and initial uranium source) to the chemical features preserved within uranium samples. In this study, a total of 81 U3O8 materials were prepared under various processing conditions. Subsequently, powder X-ray diffraction (p-XRD) and inductively coupled plasma mass spectrometry (ICP-MS) were used to identify the chemical and metallic impurities of the samples. The analysis uses regression and classification data mining trees to predict each reaction condition using both sets of impurities. The analysis shows metallic impurities are consistently chosen as explanatory variables. These signatures may assist in determining the provenance and history of unknown samples.

Joint work with Jordan Bakerman 1, Kari Sentz 2, Alison Tamasi 2, Marianne Wilkerson 2, Alyson Wilson 1
1 North Carolina State University, Raleigh, North Carolina 27695, United States
2 Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States

---

28. Learning Algorithms for Forecasting Tropical Cyclone Rapid Intensification Events
Lisa Bramer, Ph.D., Pacific Northwest National Laboratory

Machine learning algorithms can easily be implemented on streaming data. Traditionally, the application of these models to streaming data have ignored the temporal information and assumed independence of observations. However, in the streaming context, there is a continuous flow of data, rather than a static dataset, the data distribution is expected to evolve over time, and observations are seen over time in sequential order and likely are not independent. We propose the development of a framework that improves predictive analytics on data streams through the inclusion of temporal dependence in probabilistic learning models. We illustrate this idea in the context of forecasting rapid intensification events (changes in maximum wind speed of at least 30 knots within a 24 hour period) from hurricane data. A key objective of studies in this area is to develop methods to determine the probability that an RI event will occur in the next 24 hours using real time hurricane tracking data and climate model forecasts. We assess model performance using cross validation on a dataset of tropical cyclones in the Atlantic Ocean from 1982 to 2014, and results are compared to those using the current Statistical Hurricane Intensity Prediction Scheme (SHIPS).

Joint work with Bryan Stanfill (PNNL), Sarah Reehl (PNNL), Maggie Johnson (Iowa State), Petruta Caragea (Iowa State)

29. Mathematical Tools for Analysis of High Resolution, Time-Resolved 3D X-Ray Images
Talita Perciano, Lawrence Berkeley National Laboratory

An intense X-ray source--a synchrotron--can soon produce many terabytes of 3D image data per day. It is impossible to analyze all this data by hand, so analysis tools are necessary for this end. We describe a multidisciplinary work involving 3D movies from the microCT instrument of the Advanced Light Source (ALS), the National Energy Research Scientific Computing Center (NERSC), and the Center for Advanced Mathematics for Energy Research Applications (CAMERA), all located at Lawrence Berkeley National Laboratory (LBNL). The 3D movies to be featured at the event show materials changing due to their environment (heat, compression, etc.). Understanding the changes requires not just visualizing the huge amounts of data as 3D movies, but identifying and quantifying features of interest and tracking them in time. We take a multi-pronged approach to automate this problem. One is to separate the phases that compose the materials using a parallel Markov Random Field method (PMRF). Another is to use a 3D fast filtering package (F3D) and pattern matching to detect and measure specific features in the samples. We test these on synthetic data from NGCF [1], and show their application to geological as well as SiC-based composite samples.

The PMRF framework is based on graph partitioning, which allows parallel optimization and parameter estimation. The LAP algorithm [2] tailors the MRF model decomposition, such as it reduces the computational complexity drastically by applying the optimization separately for each subgraph, being exponential in the size of the largest graph clique instead.

F3D [3] is a package that provides accelerated 3D nonlinear filters, with OpenCL kernels that explores graphics cards technology. We combine F3D with machine learning, for example, template matching, to detect structures in 3D images, including fibers and cracks. Taking advantage of a model that minimizes data movement and uses intra-socket parallelism, F3D runs tens of times faster than traditional techniques.

Results show the viability and accuracy of the methods for the assessment of microCT images of material samples. Time-consuming tasks that were executed manually before are now performed automatically and efficiently. These methods have been integrated into a framework for material assessment including visualization, analysis and simulation of experiments.

Joint work with Bale, Hrishikesh - University of California Berkeley; Bethel, Wes - Lawrence Berkeley National Laboratory; Krishnan, Harinarayan - Lawrence Berkeley National Laboratory; Loring Burlen - Lawrence Berkeley National Laboratory; Mizrahi, Yariv - University of British Columbia; Parkinson, Dilworth - Lawrence Berkeley National Laboratory; Ritchie, Robert - Lawrence Berkeley National Laboratory; Sethian, James - University of California Berkeley; Ushizima, Daniela - Lawrence Berkeley National Laboratory

[1] http://xct.anu.edu.au/network_comparison
[2] Y. D. Mizrahi et. al., “Linear and parallel learning of markov random fields”, in Proceedings of International Conference on Machine Learning. June 2014, vol. 32, pp. 1-10, IMLS.
[3] D. M. Ushizima et. al., ‘structure recognition from high resolution images of ceramic composites’, IEEE International Conference on Big Data, Oct. 2014.

30. Adaptive Splines for Computer Model Emulation and Sensitivity Analysis
Devin Francom, University of California Santa Cruz

We present an emulation method based on Bayesian multivariate adaptive regression splines (BMARS) that is able to efficiently handle large numbers of model runs, large numbers of inputs, and functional output. For this emulator, we show how global sensitivity analysis can be performed without Monte Carlo error. We also present a parallel MCMC scheme to improve posterior sampling. We use this emulator to perform a sensitivity analysis of a computer model of the deformation of a protective plate used in pressure driven experiments.

Joint work with Bruno Sanso, Vera Bulaevskaya, Ana Kupresanin, Gardar Johannesson

31. Exploring and Predicting Power Usage of High Performance Computing Applications
Bruce Bugbee, NREL

As data centers and high performance computing (HPC) environments become increasingly common, detailed understanding of the energy needs of these systems is required. These facilities carry substantial energy footprints so improvements to hardware, software, or organizational policies can hold tremendous value. The National Renewable Energy Laboratory (NREL) is home to Peregrine, an HPC system capable of up to 2.26 Peta FLOPS. Apart from providing computational resources for NREL, Peregrine is equipped with Hewlett-Packard’s

integrated Lights-Out (iLO) systems. This gives researchers access to detailed power and thermal information at individual node levels. This work focuses on exploring the connections between power usage and application characteristics. This includes the nature of the application being run as well as user-requested features. Additionally, we present initial work on predicting power characteristics using application and hardware features. Preliminary success in application-based power prediction suggest future work on improving operational characteristics of Peregrine by providing more accurate power forecasts as well as better understanding of user and application behavior

Joint work with Caleb Phillips (NREL), Ryan Elmore (University of Denver), Kenny Gruchalla (NREL), Avi Purkayastha (NREL)

---

32. Probabilistic Analysis of Earthquake Occurrence in Oklahoma
Veronika Vasylkivska, ORISE, NETL

Determining the spatio-temporal characteristics of induced seismic events holds the potential for understanding why these events occur and how they propagate in time and space. Linking the seismic signature with other geologic, geographic, or anthropogenic factors could allow us to identify the causes and develop mitigation strategies that reduce the risk for the events. However, current seismic databases present both an overabundance of as-yet uncorrelated seismic events (i.e. large data) but also a paucity of geographic distributions (i.e. poor overall sampling). Thus novel combinations of geostatistical and big data approaches are needed to address this challenge.

In this work we present the results of an analysis of 17,000+ events from the Oklahoma earthquake catalog from 1970 to 2015. Oklahoma has experienced an increase in seismicity that may be associated with anthropogenic subsurface processes and is an excellent end-member comparison to the larger California database (which has a combination of natural and man-made activity). The method is based on the nearest neighbor clustering analysis of spatio-temporal distances evaluated between the particular event and all preceding events. Connecting each earthquake in the catalog to its nearest neighbor forms a single cluster which contains all examined events and is then split into smaller clusters by post processing. The analysis of the catalog containing N events requires O(N*(N-1)/2) time and space distance calculations which leads to significant computational requirements.

We show that the distribution of distances for earthquakes in Oklahoma is different from that observed for more seismically active regions (e.g. California and Japan). We believe that the observed difference can be attributed to the dominance of mostly induced events in Oklahoma. Furthermore, we investigate the properties of the largest in size earthquake clusters (often associated with the largest magnitude events) and their sensitivity to the method parameters. Sensitivity analysis makes certain that results are meaningful and do not depend on the chosen parameters.

The combination of a large dataset that must be computed over, method that creates even larger data volumes, and several input parameters that must be tested yields a computational challenge. We were able to overcome the issue by implementing a nontrivial parallelization of the process, which reduced computational time from 6-10 hours to 10-30 minutes on a desktop PC. As a bonus the resulting speed up allows us to test improvements to the underlying method and generate data that can be used for hypothetical testing of other factors which might be correlated or contributing to induced seismicity events. We hope that by analyzing and comparing the characteristics of the events from a particular cluster (and/or relevant data sets) that this method will contribute to an overarching approach for evaluating the likelihood of future seismicity.

Joint work with Nicolas Huerta, NETL; Kelly Rose, NETL; Jennifer Bauer, AECOM

---

33. Phasor Measurement Unit-based Protection Strategies for Smart Grid Cyber-Attacks
Manuel Joseph Garcia, University of Texas, Austin

Unobservable cyber attacks are undetectable by any bad data detection algorithm. In recent years, considerable attention has been paid to unobservable cyber attacks that impair the solution of the steady state estimator by compromising power injection, power flow and breaker status meters throughout the system. Such attacks can be categorized into two types. The jammer attack compromises a meter by preventing its data from reaching the state estimator. The data integrity (DI) attack compromises a meter by intelligently altering its data to false values.

Most existing countermeasures target a specific type of DI attack known as the sparse power injection data integrity (PIDI) attack, which compromises only a few power injection and power flow meters. Utilizing algorithms

initially developed to identify and protect against sparse PIDI attacks, this work provides a Phasor Measurement Unit-based countermeasure that targets general DI and jammer attacks. Special attention is paid to large-scale jammer attacks, which are labeled as being "likely" to occur due to their relative ease of deployment. This work suggests a Phasor Measurement Unit-based protection strategy to protect against large-scale jammer attacks by operating the system to "observable system states."

---

34. Climate Model Calibration Across Multiple Spatial Resolutions
Vera Bulaevskaya, Lawrence Livermore National Laboratory

Parameterizations of physical processes in climate models are highly dependent on the spatial resolution and must therefore be tuned separately for each resolution under consideration. At high spatial resolutions, objective methods for parameter tuning are computationally prohibitive. Our work has focused on calibrating parameters in the Community Atmosphere Model 5 (CAM5) at three spatial resolutions: 4, 2 and 1 degrees. Using perturbed-parameter ensembles and a statistical emulator of CAM5, we have identified input parameters at each resolution level that minimize discrepancies of key output quantities simulated by CAM5 with respect to satellite data. Our approach makes it possible to calibrate the high-resolution version of CAM5 by leveraging cheaper, low-resolution simulations and statistical models. In addition, it allows us to determine the optimal allocation of computer simulation runs among the three resolutions, i.e., one that minimizes the model's discrepancy relative to satellite data at a given computational budget. While this work deals with calibration of CAM5 model in particular, the approach is applicable to any computer model run at various degrees of spatial or temporal fidelity.

Joint work with Donald Lucas, Lawrence Livermore National Laboratory

---

35. Embedding Computer Models into a Bayesian Model for Multiple-Population Globular Clusters
David C. Stenning, Sorbonne Universite, Institut d'Astrophysique de Paris

Computer models are becoming increasingly prevalent in a variety of scientific settings; these models pose challenges because the resulting likelihood function cannot be directly evaluated. For example, astrophysicists develop computer models that predict the observed spectrum of a star as a function of input parameters such as age and chemical composition. A goal is to use such models to derive the physical properties of globular clusters — gravitationally bound collections of up to millions of stars. Recent observations from the Hubble Space Telescope provide evidence that globular clusters host multiple stellar populations, with stars belonging to the same population sharing certain physical properties. We embed computer models for stellar spectra into a statistical likelihood function that assumes a hierarchical structuring of the parameters in which physical properties are common to (i) the cluster as a whole or to (ii) individual populations within a cluster, or are unique to (iii) individual stars. A Bayesian approach is adopted for model fitting, and we devise an adaptive MCMC algorithm that greatly improves convergence relative to its precursor, non-adaptive MCMC algorithm. Our method constitutes a major advance over standard practice, which involves fitting single computer models by comparing their predictions with one or more two-dimensional projections of the data.

Joint work with David A. van Dyk, Imperial College London, London, UK, Rachel Wagner-Kaiser, Bryant Space Center, University of Florida, Gainesville, FL, Ted von Hippel, Embry-Riddle Aeronautical University, Daytona Beach, FL, Nathan Stein, The Wharton School, University of Pennsylvania, Philadelphia, PA, Elliot Robinson, Argiope Technical Solutions, FL, Ata Sarajedini, Bryant Space Center, University of Florida, Gainesville, FL

---

36. NEFDS Contamination Model Parameter Estimation of Powder Contaminated Surfaces
Timothy Gibbs, Rochester Institute of Technology

Hyperspectral signatures of powdered contaminated surfaces are challenging to characterize due to intimate mixing between materials. Most radiometric models have difficulties recreating these signatures due to non-linear interactions between particles with different physical properties. The Nonconventional Exploitation Factors Data System (NEFDS) contamination model is capable of recreating longwave hyperspectral signatures at any contamination mixture amount, but only for a very limited selection of materials currently in the database. A method has been developed to invert the NEFDS model using a non-linear least squares curve fitting algorithm to optimize model parameter estimation on emissivity measurements from various powdered materials on substrates. Emissivity data were measured using a Designs and Prototypes Fourier Transform Infrared Spectrometer for different levels of contamination. Temperature emissivity separation was performed to convert data from measured radiance to estimated surface emissivity. Emissivity curves were then input into the inverted model and parameters were estimated for each spectral curve. Performance of the inverted NEFDS

contamination model was assessed through a comparison of measured data with extrapolated model emissivity curves using estimated parameter values. Results indicate that inverting the NEFDS contamination model with measured data inputs to derive physical parameters from a real scene is a viable alternative to the limited nature of the current forward model.

Joint work with David Messinger

---

37. Enhanced Precision by Combining Sensitivity Test Data
David H. Collins, Los Alamos National Laboratory

Sensitivity testing involves applying a range of stimulus values to an experimental subject and recording a binary response; interest lies in estimating the distribution of threshold values in the subject population, where the threshold delineates response from non-response. Within the DoD/DoE complex, sensitivity testing is extensively applied to explosives, using mechanical or electrical shock as the stimulus, and detonation versus non-detonation as the response.

The choice of design points for an experiment is commonly done using one of several sequential methods. Given the data (stimulus/response pairs), analysis can be done using a direct likelihood calculation, or by binary regression (typically assuming either a logistic or Gaussian threshold distribution). Quantities to be estimated include location and scale parameters, and quantiles of the distribution.

Sensitivity testing is often used to qualify lots of explosive devices, using on the order of 20 devices per test. If the lots are assumed to be drawn from a common population, the precision of estimates can be significantly improved by analyzing data aggregated from multiple lots. We demonstrate this using simulated data, and show how to test the assumption of a homogeneous population using likelihood ratio-based confidence regions.

Similar methods can be used in cases where separate batches of devices are deliberately varied in order to test the effects of various factors on performance. In this case, the aggregated data are analyzed with covariates representing the factor levels. We present an example of an experiment designed to help determine the optimal physical design of an exploding bridgewire detonator.

Joint work with Michael S. Hamada, Los Alamos National Laboratory, Brian P. Weaver, Los Alamos National Laboratory

---

38. Uncertainty Quantification and Sensitivity Analysis in the CICE v5.1 Sea Ice Model
Jorge R. Urrego-Blanco, Los Alamos National Laboratory

Changes in the high latitude climate system have the potential to affect global climate through feedbacks with the atmosphere and connections with mid latitudes. Sea ice and climate models used to understand these changes have uncertainties that need to be characterized and quantified. In this work we characterize parametric uncertainty in Los Alamos Sea Ice model (CICE) and quantify the sensitivity of sea ice area, extent and volume with respect to uncertainty in about 40 individual model parameters. Unlike common sensitivity analyses conducted in previous studies where parameters are varied one-at-a-time, this study uses a global variance-based approach in which Sobol sequences are used to efficiently sample the full 40-dimensional parameter space. This approach requires a very large number of model evaluations, which are expensive to run. A more computationally efficient approach is implemented by training and cross-validating a surrogate (emulator) of the sea ice model with model output from 400 model runs. The emulator is used to make predictions of sea ice extent, area, and volume at several model configurations, which are then used to compute the Sobol sensitivity indices of the 40 parameters. A ranking based on the sensitivity indices indicates that model output is most sensitive to snow parameters such as conductivity and grain size, and the drainage of melt ponds. The main effects and interactions among the most influential parameters are also estimated by a non-parametric regression technique based on generalized additive models. It is recommended research to be prioritized towards more accurately determining these most influential parameters values by observational studies or by improving existing parameterizations in the sea ice model.

Joint work with Nathan Urban (Los Alamos National Laboratory), Elizabeth Hunke (Los Alamos National Laboratory)

39. Comparison of Active Subspaces and Sufficient Dimension Reduction in an MHD Model
Andrew Glaws, Colorado School of Mines

We compare the dimension reduction capabilities and interpretations of the method of active subspaces and the gradient-free approach of sufficient dimension reduction. Magnetohydrodynamic models are examined as a practical implementation of the techniques.

Joint work with Paul Constantine (Colorado School of Mines), Tim Wildey (Sandia National Labs), John Shadid (Sandia National Labs)

40. Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies
Paul Constantine, Colorado School of Mines

Scientists and engineers use computer simulations to study relationships between a physical model's input parameters and its outputs. However, thorough parameter studies---e.g., constructing response surfaces, optimizing, or averaging---are challenging, if not impossible, when the simulation is expensive and the model has several inputs. To enable studies in these instances, the engineer may attempt to reduce the dimension of the model's input parameter space. Active subspaces are part of a set of dimension reduction tools that identify important directions in the parameter space. I will describe methods for discovering a model's active subspace and propose strategies for exploiting the reduced dimension to enable otherwise infeasible parameter studies. For more info, see www.activesubspaces.org

Constantine, P.G., Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies, SIAM Spotlights Series, SIAM, Philadelphia (2015)

41. Rotation Based Compressive Sensing Method for Uncertainty Quantification with Expensive Data
Xiu Yang, Pacific Northwest National Laboratory

Compressive sensing has become a powerful tool for regression model, uncertainty quantification, data analysis, etc. The capability of obtaining an accurate approximation with a few sampling points makes this method suitable for complex systems with sparse representation. Especially when the simulation (or experiment) of the system is costly, i.e., the data is very expensive to obtain, the compressive sensing based method can be very useful. We propose an approach to quantify the uncertainty of the system with limited (expensive) data and similar idea can be applied to various different areas. We identify new random variables through linear mappings such that the representation of the quantity of interest is more sparse with new basis functions associated with the new random variables. This enhancement of sparsity improves both the efficiency and accuracy of the compressive sensing-based method. Specifically, we consider rotation-based linear mappings which are determined iteratively for Hermite polynomial expansions. We demonstrate the effectiveness of the new method with several examples.

42. Data Science Methods to Support Multi-Scale Modelling of Wellbore Integrity
Deborah Glosser, United States Department of Energy, National Energy Technology Laboratory/ORISE

As the use of subsurface reservoirs for energy production, waste disposal, and CO2 storage has grown in recent years, so have concerns about the potential for unwanted interactions with surface and near-surface receptors. Wellbores are engineered fluid flow pathways designed to facilitate the production and injection of fluids in a controlled manner. Because of the coupling between wellbore age, location (x,y,z), and materials integrity, it is critical to have a framework to evaluate the flow barrier potential of wellbores across spatial and temporal scales. Scale influences the evaluation of wellbore integrity, and modelling efforts must reflect the scale of the questions being asked. At the individual well scale, materials integrity is tied to original design and construction criteria; as well as to in situ subsurface properties such as geologic media, and resulting fluid, pressure, and flow properties. However, at the regional and meso-scale, data on specific wellbore performance is often poorly characterized for the range of subsurface conditions encountered. Therefore, the application of geoprocessing and data science techniques to evaluate spatial and temporal trends in wellbore integrity at larger scales can inform probabilistic models for whether a given wellbore will serve as a conduit for fluid flow. Here, we present two methods for assessing wellbore integrity at these two different scales. At the individual wellbore scale, we present a mathematical model that quantifies how physical forcings in contemporary foamed wellbore cement operational processes influence cement microstructure and mechanical performance. We then assess wellbore integrity at the meso-to-regional scale, and present a novel geo-processing based framework for collecting, analyzing, and communicating spatio-temporal uncertainty in wellbore datasets. Ultimately, integrating scale relationships with

spatio-temporal variations and uncertainty in data quality provides a data-driven basis for the robust modelling of these important fluid flow pathways.

Joint work with Kelly Rose (NETL), Jen Bauer (NETL), Barbara Kutchko (NETL), Dustin Crandall (NETL), Glen Benge (Benge Consulting), Tyler Ley (Oklahoma State)

---

43. Applying Dynamic Mode Decomposition to the Analysis of Climate Data
Cristina Garcia-Cardona, Los Alamos National Laboratory

In some physical systems is often the case that the dynamical behavior is governed by a set of low-order modes. Identifying such modes is of great importance for characterizing the long range dynamics of the system. Dynamic Mode Decomposition (DMD) is one of those tools that allows for detecting the coherent structures that develop in the system evolution, as well as their growth/decay rates. It can be applied to sequential and non-sequential data, generated from simulated systems or experimental measurements, and is relevant for the analysis of nonlinear systems. However, identifying the most important modes with DMD is still challenging. A kernel-based variant of DMD has been proposed as an alternative to simplify the mode detection process. In this work, we explore the utility of both DMD and kernel-based DMD for detecting modes in climate data, and discuss their merits in comparison with more traditional techniques such as Proper Orthogonal Decomposition (POD).

Joint work with Nathan Urban - Los Alamos National Laboratory

---

45. Using Bayesian Compressive Sensing to Characterize Nonlinear Chirp Signals
Matt Goodwin, Brigham Young University

Applications in remote sensing require identifying electromagnetic signals from a limited number of samples, in the presence of noise. In this setting, we consider the problem of reconstructing a nonlinear chirp signal with multiple harmonics from a few noisy, random samples. We explore the problem under the framework of compressive sensing (CS), using a Bayesian implementation known as Bayesian compressive sensing (BCS), to find a sparse representation of the signal. This process is initiated by first decomposing the sampled signal into a basis function expansion and then using the BCS algorithm to identify the potentially sparse signal, leaving us with a characterization of the signal that is robust to noise. Different basis expansions are considered, including Fourier, B-Splines, and wavelets. A comparison of basis expansions is performed based on the reconstruction accuracy with special consideration to more parsimonious representations.

Joint work with C. Shane Reese, Brigham Young University

---

46. Orthogonal Transformations for the Ensemble Kalman Filter
Humberto C. Godinez, Los Alamos National Laboratory

Data assimilation of complex physical systems is commonplace in many areas of science in order to predict phenomena of interest. In this talk we use orthogonal transformations in the ensemble Kalman filter to capture the dominant correlations within a model and reduce noise. A series of assimilation experiments are presented for a inner ring current model using real observational data. The results show a reduction in forecast error using orthogonal transformation compared with assimilation without transformation.

Joint work with Earl Lawrence -- Los Alamos National Laboratory

---

47. CITRUS: A New Framework for Detecting Materials of Interest in Spectral Imagery
Amanda Ziemann, Los Alamos National Laboratory

Remotely sensed hyperspectral imagery provides, at each pixel, a radiance spectrum with up to hundreds of distinct wavelength channels. This high-dimensional spectral information allows for pixel-level material discrimination, including applications to remotely detecting the presence of particular materials of interest within a scene. Target detection takes a spectrum (or multiple spectra) corresponding to the target material, typically laboratory-measured, field-measured, or image-derived, and uses a detection algorithm to assign a target-likelihood score to each pixel in the image. Here, we explore the tradeoff between spatial accuracy and detection accuracy by presenting the CITRUS algorithm (Cueing Image Target Regions by Unmixing Spectra), which performs target detection on local patches, or tiles. Instead of assigning a detection score to each pixel, we assign a detection score to each tile, indicating the likelihood of the target being contained within that local region; the assumption is that detecting target-containing regions will have lower false alarm rates than when detecting

target-containing pixels. Through local spectral unmixing in each tile, we identify the corresponding endmembers using the MaxD algorithm. We hypothesize that for tiles containing the target material in either full-pixel or sub-pixel abundances, the target spectrum will be within — or close to, in a Euclidean sense — the simplex described by those endmembers. The resulting detection map highlights the tiles with low target to-simplex distances, and serves as a visual cueing tool for an analyst. The ultimate application of this approach is target-region detection in broad-area spectral imagery. The methodology will be detailed here, and results will be shown on a ground-truthed hyperspectral dataset.

Joint work with James Theiler, Los Alamos National Laboratory

---

48. Model Calibration Toward Hierarchical Validation in Radiation Transport Experiments
Michael Grosskopf, Simon Fraser University

As expensive computer models become more important across areas of science and engineering, assessment of their precision and accuracy of their predictions is critical. Part of this is understanding the uncertainty in calibration parameters of the computer model and how this contributes to the overall predictive uncertainty. We present work motivated by research modeling radiation transport in high-energy-density physics experiments at the Center for Exascale Radiation Transport (CERT). There, high performance computing is used for simulating the flow of energy by radiation in physical systems. We present work addressing the unique challenges in combining the output from these models and experiments using the Kennedy-O'Hagan Bayesian framework as first steps in hierarchical validation.

Joint work with Derek Bingham, Simon Fraser University; Marv Adams, Daryl Hawkins, Aaron Holzaepfel, Texas A&M University

---

49. Optimal Condition-based Maintenance Policies under the Gamma Degradation Process
David Han, University of Texas at San Antonio

Condition-based maintenance is an effective method to reduce unexpected failures as well as the operations and maintenance costs. This work discusses the condition-based maintenance policy with optimal inspection points under the gamma degradation process. A random effect parameter is used to account for population heterogeneities and its distribution is continuously updated at each inspection epoch. The observed degradation level along with the system age is utilized for making the optimal maintenance decision, and the structure of the optimal policy is examined.

---

50. PageRank as a Microseismicity Characterization Tool: An Application to the Newberry Volcano Geothermal Site
Ana C. Aguiar, Lawrence Livermore National Laboratory

Data mining methods, such as event clustering, have often been used to explore the similarities of seismic events that comprise an earthquake sequence. Some non-standard techniques have also recently proven useful for finding similarities between seismic events. One example is PageRank (Google's initial search algorithm); a measure of connectivity for a data window, where high PageRank means high connectivity to many other data windows. PageRank, even though developed for webpage search engines, was initially used in seismology to detect low-frequency earthquakes. This application consisted in exploiting direct and indirect linkage between seismogram windows; with the advantage that linkage is computed from a sparse matrix, making calculations fast and ideal for large data sets. In seismic applications, connectivity is measured by the cross correlation of 2 time windows recorded on a common seismic station and channel. We expand on the initial application of PageRank by using it to define signal-correlation topology for micro-earthquakes in a geothermal environment. We have focused on the Newberry Volcano, within the Deschutes National Forest in Oregon, which was designated as a candidate site for the Department of Energy's Frontier Observatory for Research in Geothermal Energy (FORGE) program. This site was stimulated using high-pressure fluid injection during the fall of 2012, which generated several hundred microseismic events. Exploring the spatial and temporal development of microseismicity is key to understanding how subsurface stimulation modifies stress, fractures rock, and increases permeability. For our analysis we adapt PageRank to evaluate microseismicity during the 2012 stimulation. We then use this information to create signal families and compare PageRank families to the spatial and temporal proximity of associated earthquakes. Studying signal PageRank will potentially allow us to efficiently group earthquakes with similar physical characteristics, such as focal mechanisms and stress drop. Our ultimate goal is to determine

whether changes in the state of stress and/or changes in the generation of subsurface fracture networks can be detected using PageRank topology, as well as show how non-standard methods can be valuable in characterizing seismic sequences. This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-ABS-681113.

Joint work with Stephen C. Myers, Lawrence Livermore National Laboratory

---

51. Forensic Tool Mark Analysis: A Class and Individual Characteristic Approach
Jeremy Hadler, Iowa State University

When a striated tool such as a screw driver is used to commit a crime, often there is a tool mark left behind as evidence. It is then the job of a forensic tool mark examiner to compare this crime scene tool mark to tool marks made by a suspect's tool to determine if the crime scene mark matches any of the suspect's marks. Forensic tool mark examiners say that these striated tool marks are comprised of "class" and "individual" characteristics where class characteristics are traits common to a large number of tools such as the width of a screwdriver head and individual characteristics are traits common to a specific tool such as the striated marks left by a screwdriver. Examiners first compare marks according to their class characteristics and if they match, they continue to compare the individual characteristics. If they do not match, it is concluded the marks were not made by the same tool. Many of the algorithms being developed to replace the subjective nature of an examiners comparison ignore this class and individual characteristics idea and attempt to directly compare the marks. We have developed a procedure for comparing tool marks that accounts for the class and individual characteristics and determines two tool marks were made by the same tool only if both the class and individual characteristics of the marks align at the same location. This is accomplished by formulating a hypothesis test (common tool vs. different tool) and implementing lowess smoothing, residuals, correlation and intuitive p-values based upon the location of the largest correlation.

Joint work with Dr. Max Morris, Iowa State University

---

52. Towards Constraining Climate Sensitivity Using a Reduced-Order Climate Model
Alex Jonko, Los Alamos National Laboratory

Reducing uncertainty in equilibrium climate sensitivity, commonly defined as the surface temperature response to an instantaneous doubling of atmospheric CO2 concentrations, has been an abiding goal of the climate science community for the past decades. The range of plausible climate sensitivities, derived from ever improving generations of GCMs, however, remains essentially unchanged.

Here we present a new approach to multi-model uncertainty quantification (UQ) using a reduced-form simple climate model, tuned to reproduce relevant physics of different global climate models (GCMs). Rather than focusing on the discrete projections made by individual GCMs, our simple model allows us to smoothly interpolate between the dynamics of the multi-model ensemble by forming a continuous probability distribution over a reduced model parameter space.

We will discuss an early version of the simple model, an idealized ocean-atmosphere energy balance model (EBM). The EBM is fit to surface temperature, ocean heat content and top-of-atmosphere radiative fluxes of GCMs participating in the Coupled Model Intercomparison Project version 5 by varying several parameters, including climate sensitivity and feedback. We obtain distributions of these parameters and update the uncertainties associated with them using observations of surface temperature, ocean heat uptake and top-of-atmosphere radiative flux in a Bayesian inference framework.

Joint work with Nathan Urban

---

53. Evaluating Complex Groundwater Controls Using an Integrated Hydrologic Model of the Continental US
Laura Condon, Syracuse University

We assess complex spatial patterns in the physical controls of groundwater depth and flux using results from an integrated groundwater surface water simulation of the majority of the continental US. Findings are compared for a pre-development scenario and a simulation that reflects the total groundwater extracted over the past 100 years. Results illustrate clear multi scale behavior and regional shifts in the relative control of topography, geology and climate on water table depth and groundwater flux. In agreement with previous studies, this groundwater

simulation demonstrates relatively greater topographic control and more significant groundwater exchanges with streams in the arid west than in the humid east. We also apply a novel, k-regression algorithm to the predevelopment simulation to simultaneously identify spatial subsets of grid cells with similar relationships between explanatory variables and groundwater metrics while quantifying behavior using multiple linear regression. The combination of this statistical approach with the large-scale, high-resolution groundwater simulation allows us to quantitatively evaluate groundwater behavior across an unprecedented range of climates and physical settings. Results highlight complex spatial patterns in the relationships between groundwater and controlling variables; further demonstrating the historic difficulty in developing spatially contiguous classifications of groundwater behavior. This work highlights the potential for integrated hydrologic models to improve our ability to simplify large heterogeneous systems.

Joint work with Reed Maxwell, Colorado School of Mines

---

54. An Exposition on the Propriety of Restricted Boltzmann Machines
Andee Kaplan, Iowa State University

A restricted Boltzmann machine (RBM) is an undirected graphical model constructed for discrete random variables, specified as having two layers, a hidden and a visible layer with no conditional dependency within a layer. In recent years, RBMs have risen to prominence due to their connection to deep learning - by treating a hidden layer of one RBM as the visible layer in a second RBM a deep architecture can be created. The method is claimed to have the ability to learn very complex and rich structures in data, making these models attractive for supervised learning. However, the generative behavior of RBMs has largely been unexplored. In this presentation, we discuss the relationship between parameter specification and the prevalence of degenerate models as well as an exposition of the difficulties in fitting such highly flexible models.

Joint work with Daniel Nordman, Iowa State University, Stephen Vardeman, Iowa State University

---

55. Topological Data Modeling for Integration of Complex, Heterogeneous Information
Cliff Joslyn, Pacific Northwest National Laboratory

We report on our efforts in Topological Data Modeling (TDM), drawing on methods from topology, combinatorics, and category theory, to address the problem of multi-INT information integration in complex systems, in particular bridging between semantic and quantitative information sources. While initially proven to be tremendously valuable in traditional signal analysis (e.g. radar networks or collections of optical cameras), modern mathematical methods in areas like persistent homology, computational topology, and simplicial sheaf theory are increasingly penetrating data analytics and knowledge discovery. Our approach is grounded around three orientations:

*) Hypergraphs and abstract simplicial complexes are used as multidimensional graph structures to represent the multi-way interactions amongst information sources;

*) Categorification (a method from category theory) is used to map disparate data types into a common mathematical framework;

*) And finally mathematical sheaves are used to represent the constraints on integrated data in a multidimensional context.

As a modeling framework, TDM offers a principled, unitary approach to integrating disparate sources in complex combinations. A mathematical sheaf, in particular, is not just a mathematical structure that provides a common ground among heterogeneous data sources, it is the canonical such structure: any principled specification of the interaction between heterogenous data sources will provably recapitulate some portion of sheaf theory. Although there are no free lunches or magic bullets, in exchange for the up-front modeling of individual data sources and their pairwise interactions, what is gained is a turnkey integrated picture with inherent consistency, uncertainty quantification, and error measurement. Additionally, the topological shape of the resulting data structure is readily calcuable, potentially revealing otherwise hidden characteristics such as confounding 'data loops' which would otherwwise not be observable.

We illustrate this modeling approach in the context of a geolocation problem of a single, group target: wildlife tracking in an urban environment given a variety of information feeds, including multiple quantitative and a semantic (text) information source. We contrast the sheaf model with a corresponding stochastic systems model, and discuss the computational and modeling tradeoffs presented.

Joint work with Lauren Charles-Smith (PNNL), Vidit Nanda (U Pennsylvania), Katy Nowak (PNNL), Brenda Praggastis (PNNL), Emilie Purvine (PNNL), Michael Robinson (American University), Arun Sathanur (PNNL)

---

56. Predictive Enforcement of Hazardous Waste Regulations

Jane Zanzig, Center for Data Science and Public Policy, University of Chicago

The improper treatment and disposal of hazardous waste can lead to disastrous effects on the environment and human health. The Resource Conservation and Recovery Act (RCRA), enacted in 1976, gives the Environmental Protection Agency (EPA) the authority to regulate hazardous waste from cradle to grave. To enforce these regulations, the EPA regularly conducts inspections of facilities that handle hazardous materials; however, due to resource constraints, they can only inspect about 1300 of over 450,000 active facilities every year. The current targeting process is heuristic, based on political priorities or random selection.

We have spent the past year working with the EPA to design an unprecedented data driven approach to investigation targeting. Using reporting, monitoring, and enforcement data, we developed predictive models to identify likely violators. Validation on historical data indicates that this approach could increase inspection efficiency by an estimated 30%; however, this only provides information on inspected facilities. To validate the model's performance on novel inspections, we are conducting a field trial of approximately 50 inspections in the spring of 2016.

Joint work with Eric Potash (University of Chicago), Ben Brew, Asif Zubair (University of Southern California)

---

57. A Bayesian Hierarchical Model for Estimating Influenza Severity

Nicholas Michaud, Department of Statistics, Iowa State University

Timely monitoring and prediction of the trajectory of seasonal influenza epidemics allows hospitals and medical centers to prepare for, and provide better service to, patients with influenza. The CDC's ILINet system collects data on influenza-like illnesses from over 3,300 health care providers, and uses this data to produce accurate indicators of current influenza epidemic severity. However, ILINet indicators are typically reported at a lag of 1-2 weeks. Another source of severity data, Google Flu Trends, is calculated by aggregating Google searches for certain influenza related terms. Google Flu Trends data is provided in near-real time, but is a less direct measurement of severity than ILINet indicators, and is likely to suffer from bias. We create a hierarchical model to estimate epidemic severity for the 2014 – 2015 epidemic season which incorporates current and historical data from both ILINet and Google Flu Trends, allowing our model to benefit both from the recency of Google Flu Trends data and the accuracy of ILINet data.

To forecast for the 2014 - 2015 influenza epidemic season, we provide our model with both ILINet and GFT data from previous seasons, starting with the 2004 - 2005 epidemic season, and going through the 2013 - 2104 epidemic season. The hierarchical structure of our model allows ILINet and GFT data from previous seasons to inform epidemic severity prediction in the current season. ILINet data is modeled as being an unbiased but noisy estimate of the true, unknown influenza severity. GFT severity measurements, on the other hand, are influenced by external factors such as media coverage. These factors could consistently bias GFT severity estimates to over or under-estimate the true epidemic severity depending on the intensity of media influenza coverage in a season. To account for this potential bias in GFT data, we include a temporally correlated error term which allows over or under-predictions made by GFT data in one week to carry over into the next. Estimation is performed using the Bayesian statistical software JAGS.

We examine the increase in forecast accuracy that GFT data provides by comparing the forecasting ability of our model using both GFT and ILINet data to that of a model given only ILINet data. The two models are evaluated for their ability to predict epidemic severity multiple weeks into the future, and we find that combining up-to-date Google Flu Trends data with accurate ILINet data improves epidemic severity forecasting ability significantly.

Joint work with Jarad Niemi, Department of Statistics, Iowa State University

---

58. Materials Assurance through Orthogonal Materials Measurements

Mark H. Van Benthem, Sandia National Laboratories, Albuquerque, NM 87185-0886

Concern over counterfeit or adulterated products and materials is growing. Methods of assurance based merely on a material or product meeting a performance requirement from a specification may fail to find an undesirable material change, whether unintentional or nefarious. Likewise methods of assurance based on material

composition may be undesirable because they require destructive testing or have low probability of detecting small material changes. We approach this problem from a different perspective: measure an orthogonal set of materials properties using a series of simple, non-destructive methods to build a material 'signature', then utilize statistical models to test whether the material is what the provider purports it to be. Our method allows a range of applications from raw materials to finished parts. We will present results of this materials assurance scheme including data collection and analysis.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Joint work with Donald F. Susan, Mark A. Rodriguez, James J. M. Griego, Pin Yang, Curtis D. Mowry, David Enos and Katherine Simonson, Sandia National Laboratories

---

59. Trilinear Analysis of Multivariate Electrocardiographic Data

Mark H. Van Benthem, Sandia National Laboratories

The electrocardiograph (ECG) is an invaluable tool for detecting and diagnosing cardiac irregularities. ECG data are collected using a number of leads, typically 12, attached externally to the chest of the subject or patient. Interpretation of the electrocardiogram (also, ECG) is accomplished predominantly by a detailed examination of the sinus rhythm. In this work, we take a different approach to analyzing ECGs, employing multivariate analysis. Using data from the PhysioNet online database1, we conducted trilinear analysis of 15-lead ECG data. We will present our methods of data preprocessing, method of trilinear analysis and results of data from patients as well as control subjects.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Joint work with R. Derek West, Sandia National Laboratories, Thomas E LaBruyere, Sandia National Laboratories, Katherine M. Simonson, Sandia National Laboratories, Ross L. Hansen, Sandia National Laboratories

1. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13). PMID: 10851218; doi: 10.1161/01.CIR.101.23.e215

---

60. Data Pre-Processing for Improved Short-Term Solar Power Forecasting

Tarek Elgindy, National Renewable Energy Laboratory

This work presents a global horizontal irradiation (GHI) forecasting framework for the short term timescale. GHI and weather time-series data from previous years are used to train machine learning algorithms, which forecast point GHI values one hour into the future. We show that partitioning the training data into clusters of similar current weather and training the machine learning models separately significantly improves forecast accuracy. We also demonstrate that training the models using features of a high order polynomial which is fit to the time-series data further improves the forecasts. These algorithms are run on the seven locations in the Surface Radiation (SURFRAD) network in the United States. The forecasts are validated against observed measurements using several metrics such as Root Mean Squared Error and Mean Absolute Error.

Joint work with Bri-Mathias Hodge, Jie Zhang, Anthony R. Florita

---

61. Counter-Adversarial Community Detection: Initial Investigations

Philip Kegelmeyer, Sandia National Laboratories

The proposed poster will present initial empirical results quantifying the efficacy of various graph tampering attacks intended to undermine the utility of Louvain community detection as applied to a supply chain risk analysis.

The context for this work is that Sandia has recently initiated the Counter-Adversarial Graph Analytics (CAGA) project. CAGA's charter is to 'discover, study, and quantifiably characterize vulnerabilities in graph analysis methods induced by attack by informed, empowered adversaries, to develop remediations for those vulnerabilities, and to make the results available both in published literature and in practical, useful software'.

CAGA's initial focus is on 'evasion attacks', graph manipulations where an adversary wishes to induce a specific and local analysis failure; for instance, for a node not to be linked to its natural community. As an illustrative application of why this might matter, prior Sandia work used community detection help to assess supply chain risk. The method was to generate 'vendor graphs' (networks generated by starting at each commercial vendor's home URL and following hyperlinks from there), extract communities from those vendor graphs, and then look for properties of those communities that were indicative of 'interesting' vendors who should receive priority attention from analysts.

One such property was 'temperature'. That is, many of the individual nodes could be marked 'hot' or 'cold' based on additional sideband information, allowing the temperature of a community to be defined as the average temperature of the labeled nodes in that community. This community temperature value is particularly relevant because it proved to have the highest predictive value for indicating unlabeled vendors pre-identified by human analysts as interesting.

An adversarial vendor would thus have an interest in reducing its temperature; that is, an interest in being grouped with colder communities. Further, for this particular problem, tampering with the graph is easy as a vendor has total control of their home URL and can make adjustments as necessary with little cost.

The question for the adversary, then, is which sites to link to, and how many of them to link to? We consider an ordered list of nodes to link to be an 'false edge attack'. We have begun to generate principles for such attacks, we use those principles to instantiate specific attacks, and then investigate the efficacy of the attacks by plotting the temperature of the adversary's node as a function of the number of false edges added.

Our poster, then, will present these initial empirical assessments of a variety of attach methods, include comparison against random attacks and other baselines.

Joint work with Jeremy Wendt, Sandia National Laboratories, Ali Pinar, Sandia National Laboratories

---

62. Data Science For Fossil Energy
Slava Romanov, DOE-NETL

The unique focus of this program is on exploring the application of Big Data analytics to Fossil Energy (FE) technologies. Data science methodology does not exist for the applications and processes under investigation. This data science initiative will adapt and develop software tools so that it is possible to curate, archive and analyze large volumes of raw data, with the initial focus on fuel cell materials and advanced alloys, while they are used under extreme environments and power plant cycling conditions. Predictive materials degradation models will be validated against experimental data. Our preliminary analysis supports a hypothesis that development of the data analytics tools will help us to refine and validate the predictive models through discovery of hidden patterns and features among a large variety of the complex datasets. Our models include those with nearly excellent predictive power for the laboratory-scale experimental conditions, which still require broader validation. Certain analytical models that could be used in simulation and visualization will be designed to explicitly account for the features and relationships uncovered from Big Data. The experimental data will be gathered as part of the ongoing domain science research at NETL as well as via literature search and collaboration with external organizations. The developed analytical tools and user interfaces will allow the FE experts and technology developers to train the system and to extract hypotheses and recommended technological approaches out of the simulated and experimental data, in response to their questions and requests, through the integration of guided experimental research with computational sciences and engineering across time and length scales. It will provide the foundations of fundamental scientific understanding for advancing broad areas of science dealing with properties and behaviors of advanced materials and power plant components, which will accelerate the development of advanced technologies to support the DOE’s mission in energy and environment.

---

63. Detecting Periodic Sequences in Cyber Security Data
Matthew Price-Williams, Imperial College London

Recent approaches to cyber security involve building statistical models of computer network data. Modelling automated events and those caused by human activity separately could reduce the rate of false positives found using anomaly detection methods.

The work presented in this poster integrates a change point detection technique within a Bayesian framework to successfully identify periodic sequences of event times, where the opening event of any given sequence describes a human event generating an automated process of periodic events.

Difficulties arising from the presence of missing data are also addressed.

The methodology is validated on authentication log data from Los Alamos National Laboratory.

Joint work with Dr. Nicholas Heard, Imperial College London

---

64. Assessment and Blending of Decadal Climate Simulations.
Ana Kupresanin, Lawrence Livermore National Laboratory

The growing societal interest in climate change has generated a need to understand near term climate predictions that can assist in the decision-making affecting time scales of about a decade. Long-term projections depend strongly on estimated future emissions scenarios, with little dependence on the internal variability of the climate system at the time of initialization. Near term projections depend mostly on the inertia of the system, as well as on the evolution of the internal variability. They are less sensitive to different emission scenarios than long-term projections. The repository of the Coupled Model Intercomparison Project 5 that was the basis for the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. It contains simulations from different groups that are initialized every five years from climate states of 1960/1961 to 2005/2006. We develop hierarchical Bayesian models to blend initialized decadal simulations with long-term simulations and historical observations. Our model accounts for the variability induced by the initial conditions and the biases due to the different models. In addition, we estimate the internal variability of the climate models by studying very long control runs that are obtained with no external forcings. We consider three different climate models. Our results provide an assessment of the different simulations as well as a blended product that incorporates all different sources of variability into the posterior distribution of future climate predictions. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. LLNL-ABS-680267

Joint work with Francisco Beltran (LLNL), Gardar Johannesson(LLNL), Bruno Sanso (UCSC), Ben Santer(LLNL)

---

65. Tessera: An Open Source Environment for Deep Analysis of Large, Complex Data
Landon Sego, Pacific Northwest National Laboratory

Tessera (http://tessera.io) is a set of open-source packages that extend the R language to compute on large datasets via the divide and recombine paradigm. This approach involves dividing large datasets into meaningful subsets (e.g. temporally, geographically, by subject, by entity, etc.) and then computing on those subsets in parallel. Results from the computations can then be combined to achieve a particular purpose: a statistical calculation, a visualization, or another round of computation. For large data, Tessera utilizes Hadoop to manage the distributed data and computations. For smaller data or for testing prior to scaling, Tessera can be conveniently used without Hadoop in a workstation environment. Trelliscope, a component of Tessera, is a web-based interface that allows users to create customized visualizations that are replicated for each subset of the data. It makes it easy for the user to sort and filter the visualization panels by means of user-defined metrics that are calculated for each subset. Trelliscope is capable of processing hundreds of thousands, even millions, of plots. We have found it to be an invaluable tool for rapid, flexible exploratory data analysis of large and complex datasets in a wide variety of domains.

Joint work with Ryan Hafen, Hafen Consulting, LLC, Amanda White, Pacific Northwest National Laboratory

---

66. Normalized Graph Edit Distance
Luke Rodriguez, Pacific Northwest National Laboratory

Accurate and meaningful comparison of large complex networks is an inherently difficult and computationally intensive task. Most commonly available comparison schemes focus on traits of a graph such as circumference, girth, or subgraph containment, and are often heavily dependent on the relative sizes of the graphs being compared. We present the Normalized Graph Edit Distance (NGED), a new comparison metric that builds upon the basic concepts of standard Graph Edit Distance (GED), but normalizes comparisons across all graph sizes to be on the same scale. NGED can be extended to apply to attributed graphs such as those representing complex

shipping networks. Through a comparison of this new metric to GED on shipping data, we demonstrate the effectiveness of NGED in highlighting meaningful relationships between networks.

Joint work with Emilie Purvine

---

67. Uncertainty Quantification of 3D Geophysical Models

Carene Larmat, EES-17, Los Alamos National Laboratory

Seismic waves are unique geophysical observables that combine multiple wavelengths and large travelling distances, allowing subsurface image as they bring back to the surface precious information about the structure of the Earth's interior. Tomography methods are used to infer elastic properties in depth from measurements made at the surface such as travel times, dispersion or complete waveforms. 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 such as monitoring nuclear test treaties. We believe that the current computational power makes it now possible to tackle this important scientific question using the best practices from both statistical sciences and geophysical modeling.

The goal of this paper is to identify cross-disciplinary opportunities. We first review the state-of-the art in 3D earth modeling pointing to methods that are emerging to address the question of uncertainty quantifiation (UQ) in earth sciences: stochastic methods such as Monte-Carlo Markov Chain, combined with Spectral Element Method (SEM) for full wave modeling, and adjoint methods to compute kernels and Hessian. We then present results from a study we performed to validate newly developed 3D earth models for the western USA (dynamic north America (DNA) models). Two DNA models generated with different imaging techniques were compared; one developed using ray theory, and one using finite frequency theory. This comparison addressed a current debate in seismology regarding the adequacy of ray theory for seismic imaging. We use LANL high-performance computing resources to compute a set of synthetic seismograms using the SEM. The number of measurements made on those synthetics was large enough to conduct a statistical assessment. Results show that differences for synthetics computed with the two models can be seen for small periods (< 15s) and are mostly present for large negative anomalies such as the Yellowstone plume.

Joint work with Monica Maceira, EES-17, Los Alamos National Laboratory; Rob Porritt, University of Southern California; Dale Anderson, EES-17, Los Alamos National Laboratory; David Higdon, Social Decision Analytics Laboratory,Virginia Tech University

---

68. Analysis of Community Evolution in Networks

Goran Konjevod, Lawrence Livermore National Laboratory

We are interested in analyzing change in dynamic networks. More precisely, we consider the changes in the activity distribution within the network, in terms link existence and intensity link weight. Detecting change in univariate metrics reduces to statistical process control. Detecting change in larger-scale structures is more challenging and less well understood. We study the problem of classifying nodes by their role in the process underlying the network, as well as detecting changes in networkstructure.

Using a dataset covering all taxi trips in New York City since 2009, we investigate the evolution of an ensemble of networks under different spatio-temporal resolutions. We identify the community structure by fitting a weighted stochastic block model. We study node ranking and clustering methods, their ability to capture the rhythm of life in the Big Apple, and their potential usefulness in highlighting changes in the underlying network structure.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

Joint work with Giuliana Pallotta, LLNL

69. Probabilistic Forecasting, Stochastic Optimization, and The Progressive Hedging Algorithm
Bismark Singh, The University of Texas at Austin, Sandia National Laboratories

Integrating renewable energy into the US power grid has been a major governmental and industrial focus over the past decade, and the penetration levels of renewables are projected to significantly increase over the next two decades. Renewable energy causes major problems for power grid operations, due to both variability and uncertainty in power production. Numerical weather prediction models and historical data can be used to forecast such quantities, but ultimately with limited precision. Consequently, models and data must be leveraged to create probabilistic and non-parametric forecasts of the production levels of renewables. From such models, discrete scenarios can be sampled, which can then be provided as input to stochastic variants of existing operations models, such as day-ahead and reliability unit commitment, and economic dispatch.

Typically, a large number of scenario samples are required to obtain robust solutions for power grid operations. The resulting stochastic optimization models are known to be computationally challenging, such that these models are not presently solved exactly in practice. However, recent advances in decomposition algorithms and parallel computing leverage the fact that the individual scenario problems may be easy to solve, and can be iteratively and collectively manipulated to obtain solutions to large-scale stochastic operations problems for the power grid. The progressive hedging algorithm, originally developed by Rockafellar and Wets, is one such algorithm. Past studies have considered applications of the progressive hedging algorithm in stochastic unit commitment models. Here, we present recent advances that further the effectiveness of progressive hedging for solving industrial-scale stochastic unit commitment models, and discuss ideas for subsequent research. Ultimately, our goal is to move the combination of (i) model and data-based probabilistic forecasting, and (ii) stochastic operations models, into practice.

Joint work with Jean-Paul Watson, Sandia National Laboratories

70. Detecting Anomalous Activity in Dynamic Cybersecurity Networks
Nathan Lemons, Los Alamos National Laboratory

Motivated by the problem of detecting anomalous subgraphs in a graph of computer network connections we consider the following problem: given a time series observation of a graph G=(G_1,G_2,…,G_m), when can we determine if G was generated under the temporal graph null model H_0 or under the alternative H_1? We consider simple Markovian random graph models where the graph at time t depends only on the graph at time t-1. We investigate and compare various tests, from both an algorithmic and statistical point of view, to determine when the models are distinguishable.

Joint work with Aric Hagberg (Los Alamos National Laboratory), Sidhant Misra (Los Alamos National Laboratory)

71. Application of The Full Bayesian Significance Test to Model Selection under Informative Sampling
Anna Sikov, Universidade de Sao Paulo

Survey sampling distinguishes the cases of non-informative and in-formative sampling. In the informative sampling case, the sampling scheme is explicitly or implicitly associated with the variable under investigation. Consequently, the distribution of this variable holding in the sample tends to be different from the distribution holding in the population. In this research, we propose an approach for identification of the model underlining the sample selection mechanism, and the model, holding in the population, based on the sample measurements. We show, how the weighted estimation equations, based on the sample measurements can be incorporated into a Bayesian approach, and how the The Full Bayesian Significance Test (FBST) can be implemented as a model identification tool. We illustrate the efficiency of the proposed methodology by a simulation study.

Joint work with Julio M. Stern (Universidade de Sao Paulo)

72. A Randomized Algorithm for Zonotope Approximation
Kerrek Stinson, Colorado School of Mines

We propose a randomized algorithm to enumerate the vertices of a zonotope. As the algorithm may not return all vertices of the zonotope, we provide a probabalistic bound on the Hausdorff distance between the convex hull formed by the algorithm's returned vertices and the zonotope.

73. Modern Statistics for Pulse Shape Discrimination
Kristin P. Lennox, Lawrence Livermore National Laboratory

A key challenge in fast neutron detection is the neutron/gamma discrimination problem, where both gamma rays and neutrons are detected and must be distinguished from one another. Data takes the form of a pulse shape coming from a scintillator, which historically could only be sampled at very low resolution. Consequently, neutron/gamma discrimination was performed in two-dimensional summary spaces rather than full pulse space. A key advance in the field has been the increasing use of high resolution digitizers to more comprehensively sample full pulse shapes; however, little has been done to leverage the additional information for the neutron/gamma discrimination problem. Our work focuses on applying modern statistical techniques, particularly functional data analysis, to the neutron/gamma discrimination problem with promising early results.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Joint work with Andrew Glenn, Ron Wurtz

74. Data Mining to Assess Power Output in Non-Uniform Onshore Wind Farms
Andrea Staid, Sandia National Labs

Estimates of wind farm power production are needed for many types of planning problems, both short- and long-term. There is significant focus on characterizing and predicting wind speeds, but the subsequent conversion to power production is often overly simplified at the farm level. An accurate prediction of power output relies on detailed knowledge of wind conditions at each turbine, and this information is often difficult to estimate in real farms due to limited measurement and complex turbine wake interactions. For this reason, heuristics or simplified models are used. For long-term planning decisions, the discrepancies introduced by these simplifications can be significant.

In this research, I use data from two onshore wind farms to compare data analytic techniques for power assessment to several other methods, including the Jensen model, a standard baseline wake decay model that is heavily used in academia and industry. Wake decay models have been validated primarily in offshore and near-shore settings, although they are still used in practice for onshore applications. I present a comparison of the validity of wake models in onshore settings to statistical methods that do not explicitly model turbine wakes and interactions.

Wake interactions result in reduced wind speed, and therefore reduced power output, at downstream turbines, and this results in significant power losses in operational wind farms. Wake models such as the Jensen model are appropriate in cases where turbines in a farm are spaced closely enough so that the wakes from leading turbines do interact with downstream turbines. The exact distance at which a wake ceases to impact downstream turbines depends on the turbine size, the amount of turbulence in the surrounding flow, and any characteristics interfering with the flow, such as terrain features. Based on the two farms analyzed here, it is reasonable to expect wake effects to be present in these two farms for spacings of 15 diameters, and the effects are expected to be substantial for spacings less than 10 diameters. These spacings affect the majority of both farms for a wide range of wind directions.

I present statistical models that are trained and tested on an individual farm as well as models that are trained on one farm and tested on the other farm, for a full independent comparison. The specific turbine layout should drive the power estimate of each farm, but I show that even the models without specific farm data outperform the alternative methods, including the Jensen wake model. Because of the simplistic assumption of steady-state input flow to a wind farm, the Jensen model fails to capture the complexities of actual wind flow through a farm and therefore cannot provide accurate power predictions. Alternatives to the Jensen model, including statistical models trained with similar settings, show more promise for accurate resource assessment and farm power predictions to be used in planning decisions.

Joint work with Claire VerHulst, U.S. Military Academy, Seth D. Guikema, University of Michigan

75. High-Dimensional Intrinsic Interpolation using Gaussian Process Regression
Charanraj Thimmisetty, University of Southern California

As natural resources become scarcer, their recovery becomes more crucially dependent on illuminating the subsurface. While seismic data acquisition, processing and interpretation is well established, the value of information its gleans about the subsurface continues to improve as function of sensor and computational technologies. New modalities of subsurface exploration, however, solicit coupled interactions, thus complicating the task of signal deconvolution and interpretation. The interdependence between these different observables carries signature of multiscale spatial and temporal scales. Recent data analytics techniques permit such analysis by first considering an ambient Euclidean space containing the data, and then discovering a low dimensional manifolds embedded in this space. Based on these techniques, we introduce a regression procedure for intrinsic variables constrained onto a manifold embedded in an ambient space. The procedure is meant to sharpen high-dimensional interpolation by introducing constraints delineated from within the data being interpolated. Our method augments manifold learning procedures with a Gaussian process regression (GPR). We first identify, using diffusion maps, a low dimensional manifold embedded in an ambient high dimensional space associated with data. We rely on the diffusion distance associated with this construction to define proximity between points on the manifold. This distance is then used to compute the correlation structure of a Gaussian process that describes the dependence of quantities of interest in the high dimensional ambient space. The proposed method is applicable to arbitrarily high dimensional data and is demonstrated in the context of characterizing the permeability of rock formations from petrophysics data. For the dataset considered in this work, clusters obtained from the diffusion maps are different from the clusters obtained from principle component (PC) analysis space. Spread of the data points in each cluster in the diffusion space, is minimum compared to the spread in the PC space. While the cluster arrangement, in the PC space is similar to the spatial locations even after data normalization, diffusion maps gave different clusters. Considerable improvement in the predictions was observed with the proposed approach

Joint work with Roger Ghanem, University of Southern California, Los Angeles, California,USA; Nicolas Leseur, Saudi Aramco, Dhahran, Kingdom of Saudi Arabia.

76. A Stochastic Approach to Computing the Impulse Response of a High-energy X-ray Imaging System
Aaron Luttman, National Security Technologies, LLC

It is essential in quantitative radiography to deconvolve the imaging system response out of measured imagery in order to compute absolutely correct values of object properties like densities or feature locations. For high-energy X-ray systems, as are commonly used in the U.S. Department of Energy research complex, it is necessary to experimentally determine the system response (Point Spread Function or PSF) by measuring calibration imagery and computing the PSF from the calibration. In this work we present a hierarchical Bayesian model for computing the PSF of an X-ray system from a particular calibration object, a partially collapsed Gibbs sampler for efficiently computing samples from the posterior, and results from a real high-energy system at the U.S. Department of Energy's Nevada National Security Site.

Joint work with Kevin Joyce (University of Montana), Johnathan Bardsley (University of Montana)

77. Federating Cyber and Physical Models for Event-Driven Situational Awareness
Eric Stephan, Mark Rice, Pacific Northwest National Laboratory

This poster describes a novel method to improve electric power system monitoring and control software application interoperability. This method employs the concept of federation, which is defined as the use of existing models that represent aspects of a system in specific domains (such as physical and cyber security domains) and building interface to link all of domain models. Federation seeks to build on existing bodies of work. Some examples include the Common Information Models (CIM) maintained by the International Electrotechnical Commission Technical Committee 57 (IEC TC 57) for the electric power industry. Another relevant model is the CIM maintained by the Distributed Management Task Force (DMTF); this CIM defines a representation of the managed elements in an Information Technology (IT) environment. The power system is an example of a cyber-physical system, where the cyber systems, consisting of computing infrastructure such as networks and devices, play a critical role in the operation of the underlying physical electricity delivery system. Measurements from remote field devices are relayed to control centers through computer networks, and the data is processed to determine suitable control actions. Control decisions are then relayed back to field devices. It has been observed that threat actors may be able to successfully compromise this cyber layer in order to impact power system

operation. Therefore, future control center applications must be wary of potentially compromised measurements coming from field devices. In order to ensure the integrity of the field measurements, these applications could make use of compromise indicators from alternate sources of information such as cyber security. Thus, modern control applications may require access to data from sources that are not defined in the local information model. In such cases, software application interfaces will require integration of data objects from cross-domain data models. When incorporating or federating different domains, it is important to have subject matter experts work together, recognizing that not everyone has the same knowledge, responsibilities, focus, or skill set.

Joint work with Siddharth Sridhar, Ronald Pawlowski, Sumit Purohit

---

78. Optimal Bayesian Experimental Design for Subsurface Characterization in Contaminated Areas
Panagiotis Tsilifis, University of Southern California

Experimental design is crucial for inference where limitations in the data collection procedure are present due to cost or other restrictions. Optimal experimental designs determine parameters that in some appropriate sense make the data the most informative possible. In a Bayesian setting this is translated to updating to the best possible posterior. Information theoretic arguments have led to the formation of the expected information gain as a design criterion. This can be evaluated mainly by Monte Carlo sampling and maximized by using stochastic approximation methods, both known for being computationally expensive tasks. We propose a framework where a lower bound of the expected information gain is used as an alternative design criterion. In addition to alleviating the computational burden, this also addresses issues concerning estimation bias. The problem of permeability inference in a large contaminated area is used to demonstrate the validity of our approach where we employ the massively parallel version of the multiphase multicomponent simulator TOUGH2 to simulate contaminant transport and a Polynomial Chaos approximation of the forward model that further accelerates the objective function evaluations. The proposed methodology is demonstrated to a setting where field measurements are available.

Joint work with Roger Ghanem, Dept. of Civil Engineering, University of Southern California

---

79. Modeling Material Stress Using Integrals of Gaussian Markov Random Fields
Peter W. Marcy, Los Alamos National Laboratory

Material scientists are interested in the variability of stress conditions as compressive forces are applied throughout a volume. Sophisticated computer codes are used to simulate von Mises stress-fields, and it is often observed that the internal grain boundary structure is important. We describe the stress-field within a realized cube of tantalum (having tens of grains and many thousands of computational elements) with a model featuring integrals of GMRFs on second- and third-order grain boundaries.

---

80. Multimodal Data Integration under Uncertainty
David Stracuzzi, Sandia National Laboratories

Recent years have witnessed an explosion in the availability and application of remote sensor data to detecting and classifying objects and events. Applications range from scientific modeling, such as climate change analysis, to surveillance for intelligence purposes. With the increase in data availability comes new opportunities to improve analytic results and reduce uncertainty by combining the data from multiple sensors that observe the same region. In this work, we develop methods for estimating the uncertainty associated with the integration of data from multiple sensor modalities. For our purposes, multimodal data refers to data from multiple sources (not limited to imagery), at varying resolutions and geographical and temporal coverage. Integration in this context refers to combining data sources to produce a more complete view of the state and activity in the sensed area than any one source could produce alone (data fusion is a subset of integration). Our approach produces probabilistic information about the geographical location and semantic labels associated with detected objects in the sensor data. We present preliminary results from the combination of three different sensors, optical, LiDAR, and polSAR, and look ahead toward using the uncertainty distributions to assess the relative contribution of each source to the final result.

Joint work with David Robinson, Sandia National Laboratories; Matthew Peterson, Sandia National Laboratories; Stephen Dauphin, Sandia National Laboratories; Max Chen, Sandia National Laboratories; Robert Riley, Sandia National Laboratories

81. Empirical Mode Decomposition and Missing Data
D. Alexandra Williams, Brigham Young University

Empirical Mode Decomposition (EMD) like the Fourier transformation can break a time series into various frequency signals, called intrinsic mode functions (IMF's). However, in contrast to the Fourier transformation, the EMD is data driven and therefore allows for non-constant frequencies in each of its IMF's. This is useful for many environmental and other real-life applications; for example, El Nino occurs irregularly every two to seven years, rather than a regular, say, five-year occurrence. However, EMD cannot handle missing data. In this poster, we examine various simple data imputation methods and their ability to recreate the IMF's of a given time series. We find that using regression to predict the missing portions of each IMF accurately returns the true time series. The EMD on this completed time series can then be used to detect signal anomalies.

82. Modeling the Energy Output of Solar Panels with the Use of Finite Regression Mixtures
Igor Melnykov, Colorado State University - Pueblo

Finite mixture models have been widely used in modeling data sets that have complicated structure and can not be easily described by one of basic probability distributions. In our study, we apply a mixture of regression models to predict the energy output of solar panels based on forecast weather conditions. The predictor variables taken into account included the number of hours of daylight, temperature, sky index, humidity, precipitation amount, and other atmospheric and environmental variables. The mixture of regressions was investigated for the sets of solar panels at two locations, in Pueblo, Colorado and Norfolk, Virginia. The comparison with the standard multiple regression techniques confirmed the high versatility of the mixture modeling approach.

Joint work with Huseyin Sarper, Old Dominion University

83. Real-Time Detection of In-flight Damage
Brenton Blair, Lawrence Livermore National Laboratory

When there is damage to an aircraft, it is critical to be able to quickly detect and diagnose the problem so that the pilot can attempt to maintain control of the aircraft and land it safely. We develop methodology for real-time classification of flight trajectories to be able to distinguish between an undamaged aircraft and five different damage scenarios. Principal components analysis allows a lower-dimensional representation of multi-dimensional trajectory information in time. Random Forests provide a computationally efficient approach with sufficient accuracy to be able to detect and classify the different scenarios in real-time. We demonstrate our approach by classifying realizations of a 45 degree bank angle generated from the Generic Transport flight simulator in collaboration with NASA.

Joint work with Herbie K.H. Lee (UC Santa Cruz) Misty Davies (NASA Ames Research Center)

84. Active Subspaces for Sensitivity Analysis
Paul Diaz, Colorado School of Mines

Active subspaces are an emerging set of tools for parameter studies. We construct global sensitivity metrics from active subspaces and compute the metrics for a variety of mathematical models.

Joint work with Dr. Paul G. Constantine
Constantine, Paul G. 'Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies. SIAM.' (2015).
Diaz, Paul, and Paul G. Constantine. 'Global sensitivity metrics from active subspaces.' arXiv preprint arXiv:1510.04361 (2015).

85. pbdR: Data Analytics at Scale
George Ostrouchov, Oak Ridge National Laboratory

The pbdR project is a set of R packages for large scale, distributed computing and profiling for data analytics. High performance, high-level interfaces to MPI, ScaLAPACK, ADIOS, NetCDF4, and more. While these libraries shine brightest on large distributed platforms, they also work rather well on small clusters and often, surprisingly, even on a laptop with only two cores. http://pbdr.org

86. Echo Analysis Management Software
Stuart Taylor, Los Alamos National Laboratory, AET-1

Echo™ is a Matlab-based software package designed for robust and scalable analysis of complex data workflows. An alternative to tedious, error-prone conventional processes, Echo is based on three transformative principles for data analysis: self-describing data, name-based indexing, and dynamic resource allocation. The software takes an object-oriented approach to data analysis, intimately connecting measurement data with associated metadata. Echo operations in an analysis workflow automatically track and merge metadata and computation parameters to provide a complete history of the process used to generate final results, while automated figure and report generation tools eliminate the potential to mislabel those results. History reporting and visualization methods provide straightforward auditability of analysis processes. Furthermore, name-based indexing on metadata greatly improves code readability for analyst collaboration and reduces opportunities for errors to occur. Echo efficiently manages large data sets using a framework that seamlessly allocates resources such that only the necessary computations to produce a given result are executed. Echo provides a versatile and extensible framework, allowing advanced users to add their own tools and data classes tailored to their own specific needs. Applying these transformative principles and powerful features, Echo greatly improves analyst efficiency and quality of results in many application areas.

Joint work with Dustin Harvey (LANL AET-1), Eric Flynn (LANL NSEC), Colin Haynes (LANL W-14), John Heit (LANL W-15)

87. Intrahost HIV Evolution Requires Infinite-Variance Models
Timothy Wallstrom, Los Alamos National Laboratory

The classical model of population genetics makes a finite variance assumption, which is analogous to the assumption that is used to derive the central limit theorem. If we relax this assumption in the central limit theorem, we obtain a wider class of limits, namely the stable distributions. If we relax the corresponding assumption in population genetics, we obtain a wider class of genetic models, which are qualitatively different from the classical models. But which model correctly describes nature?

Using genetic sequence data, we show that the more general models are a better fit to intrahost HIV dynamics, and that the classical model, which is almost universally used to model such dynamics, is rejected by the data. This suggests that the underlying source of stochasticity has been misidentified in the classical models of population genetics, at least for some populations.

Joint work with Tanmoy Bhattacharya (LANL), Jon Wilkins (Ronin Institute), Will Fischer (LANL)

88. Bayesian Networks with Prior Knowledge for Malware Phylogenetics
Diane Oyen, ISR-3, LOS ALAMOS NATIONAL LABORATORY

Malware phylogenetics help cybersecurity experts to quickly understand a new malware sample by placing the new sample in the context of similar samples that have been previously reverse engineered. Recently, researchers have begun using malware code as data to infer directed acyclic graphs (DAG) that model the evolutionary relationships among samples of malware. A DAG is the ideal model for a phylogenetic graph because it includes the merges and branches that are often present in malware evolution. We present a Bayesian network discovery algorithm for learning a DAG via statistical inference of conditional dependencies from observed data with an informative prior on the partial ordering of variables. Our approach leverages the information on edge direction that a human can provide and the edge presence inference that data can provide. We give an efficient implementation of the partial-order prior in a Bayesian structure discovery learning algorithm, as well as a related structure prior, showing that both priors meet the local modularity requirement necessary for the efficient Bayesian discovery algorithm. We apply our algorithm to learn phylogenetic graphs on three malicious families and two benign families where the ground truth is known; and show that compared to competing algorithms, our algorithm more accurately identifies directed edges.

Joint work with Blake Anderson, Cisco Systems, Christine Anderson-Cook, LANL

89. Analysis of Streaming Data Using Compression Algorithm Metadata
Dr. Jeremy Teuton, PNNL

OBJECTIVE: We test the ability of algorithms used in data compression to call and classify events in steaming data in real or near real time. We use the combination of algorithmic outputs from compression metadata to give a

first classification of event for an analyst or automated tool to use in real time in a work flow or guide following analysis. As the stream continues, events can be counted with less effort, so that analysis becomes more automated, thus allowing the analyst to dig deeper without fear of missing important events.

APPROACH: Video compression uses multiple calculations comparing current state to previous state to identify changes. In this work, the value of these algorithms to speak to specific changes will be leveraged with native image/video data and with non-video data sources processed as video. Statistics on what percentage of an image moves, what fraction moves in the same direction, contiguous shapes or isolated blocks, novel vs. shifted pixels, etc. will be used in data analysis to identify and potentially classify events in streaming data.

ACHIEVEMENTS: Matched or exceeded performance of native video scene selection software without requiring training or tuning common to those tools

Demonstrated robustness to poor data quality

Automated detection of first point of failure in lithium battery (electron microscopy)

Event detection in NMR data

IMPACT: Production of a new, flexible modular tool for streaming data analysis will support the analysis of streaming data. Gaining a solid understanding of current strengths and weaknesses of this type of analysis compared to the state-of-the-art rule-based video and streaming data analysis will define both its best use cases and areas for further research and improvement.

This work will provide a strong basis for further development of the toolset and identification or creation of optimal algorithms and should also provide a solid research tool to determine cost/benefit tradeoffs in computational power and required training.

FUTURE WORK: Develop this concept further and increase its value to streaming data analysis;

Research the computational costs of each algorithm so that computational requirements can be set in advance of bringing data streams online and/or to triage selected algorithms back to remain within computational limits

Identify the effect of data quality on analysis

Determine cross-training benefits and pitfalls, which lessons about the outputs of specific algorithms are data set-specific and which may be used to shorten training time on disparate data sets, and identify if relationships between specific algorithms outputs hold true across data sets

Expand to novel data types, novel algorithms (text compression, writing level); enable faster analysis for current large data challenges including electron microscopy, crystallography analysis, and compressive sensing

---

90. Survival Model Selection with Missing Data and Correlated Covariates

Sydeaka Watson, PhD, The University of Chicago, Biostatistics Laboratory

A novel combination of existing methods was used to develop a survival prediction equation for pulmonary arterial hypertension patients awaiting lung transplantation. The Scientific Registry of Transplant Recipients (SRTR) dataset featured censored survival times, missing covariate data, and a large number of highly correlated candidate predictor variables. Penalized weighted least squares regression was repeatedly applied to bootstrap resamples of multiply imputed data, yielding a parsimonious model that satisfied internal validation criteria of clinical interest. Simulation studies under various degrees of predictor variable missing-ness, survival time censoring, effect size, and proportion of variables unrelated to survival have shown that this method tends to accurately recover the true list of Cox regression predictor variables.

---

91. Guided Search for Organic Photovoltaic Materials Using Predictive Data Modeling

Caleb Phillips, National Renewable Energy Laboratory

In this work we explore the use of data mining, informatics, and machine learning as potential methods for accelerated machine-assisted search for new organic electronic materials for organic photovoltaics (OPV), i.e., plastic solar cells [1,2]. Materials discovery for OPV has been driven by the tuning of electronic energy levels to maximize optical absorption and charge and power generation. These properties are related to the materials' optical band gap, the overlap of the absorption spectrum with the solar spectrum (spectral overlap) and the highest occupied molecular orbital (HOMO). The goal of this work is to perform inexpensive computations to

screen a large number of random candidate molecular structures, generated combinatorially from discrete molecular building blocks, for their electronic properties. The full set is then to be pruned to apply higher-level electronic structure methods to those most likely to have desirable properties.

We model our efforts on “big data” cheminformatic approaches to structure search favored in the pharmaceutical literature (e.g., [3]). Our assistance system profits from information gained during past computational simulations and stored in an extensive informatics system, allowing detailed knowledge to be built about relationships between structure properties and performance metrics. In this way, nothing is wasted and each simulated result improves the accuracy and throughput of subsequent searches. In order to maintain generalizability to other problems and other scenarios, we focus our analysis on common, well-understood cheminformatic features and molecular fingerprints and utilize “off the shelf” (OTS) open source informatics libraries.

We evaluate three models of increasing complexity: multiple linear regression (LR), multiple adaptive regression splines (MARS), and support vector regression (SVR).

To evaluate the performance of the various modeling techniques, we perform a standard cross-validation, where 25% of the data are withheld for performance testing while models are trained on the remaining 75%. This train/test process is repeated with random subsets 10 times.

Our initial results show that a polynomial SVR model is most successful at predicting semiconductor band gap and HOMO (0.18 to 0.24 EV RMSE), while the simpler models better predict spectral overlap, a measure of applicability for photovoltaic applications. This work demonstrates that predictive data modeling can accelerate the process of discovery of desirable materials while minimizing costly numerical simulations.

Joint work with Ross E. Larsen (NREL, CSC), Nikos Kopidakis (NREL, CMC), Kristin Munch (NREL, CSC)

[1] Kippelen, Bernard, and Jean-Luc Bredas. 'Organic photovoltaics.' Energy & Environmental Science 2.3 (2009): 251-261.
[2] Roberto Olivares-Amaya, Carlos Amador-Bedolla, Johannes Hachmann, Sule Athan-Evrenk, Roel S. Sanchez-Carrera, Leslie Vogt, Alan Aspuru Guzik. Accellerted computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. Energy Environ. Sci., 2011, 4, 4849.
[3] Jurgen Bajorath. Cheminformatics: Concepts, Methods, and Tools for Drug Discovery. Humana Press Inc. 2004.

---

92. PlanetSense: The Efficacy of Applying Open and Crowd Source Data in Urban Dynamics Research

Gautam Thakur, Oak Ridge National Laboratory

Geospatial intelligence has traditionally relied on the use of archived and unvarying data for planning and exploration purposes. In consequence, the tools and methods that are architected to provide insight and generate projections only rely on such datasets. Albeit, if this approach has proven effective in several cases, such as land use identification and route mapping, it has severely restricted the ability of researchers to inculcate current information in their work. This approach is inadequate in scenarios requiring real-time information to act and to adjust in ever changing dynamic environments, such as evacuation and rescue missions. In this work, we propose PlanetSense, a platform for geospatial intelligence that is built to harness the existing power of archived data and add to that, the dynamics of real-time streams, seamlessly integrated with sophisticated data mining algorithms and analytics tools for generating operational intelligence on the fly. The platform has four main components — i) GeoData Cloud — a data architecture for storing and managing disparate datasets; ii) Mechanism to harvest real-time streaming data; iii) Data analytics framework; iv) Presentation and visualization through web interface and RESTful services. Using two case studies, we underpin the necessity of our platform in modeling ambient population and building occupancy at scale.

Joint work with Gautam Thakur, Kevin Sparks, Robert Stewart, Marie Urban, Shweta Gupte, Budhendra Bhaduri, Dalton Lunga, John Quigley

---

93. Terrain Classification Techniques for High-Resolution Dual-Polarization Synthetic Aperture Radar Images

Derek West, Sandia National Laboratories

Synthetic Aperture Radar (SAR) is a remote sensing modality that provides its own illumination on a scene being imaged and measures the backscattered response. Single polarization SAR measures the backscattered responses of a scene with a fixed transmit and receive electric field polarization. The images formed from a single polarization SAR sensor are monochromatic and display the intensity of the backscattered returns from the scene.

A polarimetric SAR (PolSAR) sensor measures a scene with multiple transmit and receive electric field orientations. The data set collected by a PolSAR sensor contains not only the intensity of the backscattered returns from the scene from each polarization orientation, but it can also be processed in such a way as to give an estimate of the underlying scattering physics of the reflectors in the scene; thus, a PolSAR sensor collects a much richer data set than a single polarization SAR sensor. Polychromatic images formed from PolSAR data can be used to display the intensity of the backscattered return and an estimate of the canonical scattering mechanism type (such as surface, vegetation, etc.).

A variety of terrain classification techniques have been proposed for PolSAR data; some machine learning (ML) approaches to terrain classification of PolSAR data utilize deep belief networks and support vector machines. While a recent survey of some ML algorithms on fully-polarimetric SAR data and a comparison of classification accuracy between dual-polarization and fully-polarimetric SAR images have been conducted, not much has been done in the way of a comparative study strictly for high-resolution dual-polarization SAR data. We propose to do a comparative study based on determining how some selected ML techniques compare to a well-established frequentist-based one-class classifier for classification of terrain in high-resolution dual-polarization SAR imagery.

Joint work with Thomas LaBruyere (Sandia National Laboratories), Katherine Simonson (Sandia National Laboratories), Ross Hansen (Sandia National Laboratories), Mark Van Benthem (Sandia National Laboratories)

---

94. Application of Data Analytics to Additive Manufacturing
Sarah Powers, Oak Ridge National Laboratory

Recent advances in additive manufacturing have led to many success stories of large 3D printed objects (e.g., the Shelby Cobra car) and leave the industry poised for rapid growth.

One of the many challenges is the certification process, which currently must be done for each part. During the build progression, many parameters have the potential to influence the formation of defects as well as the final geometry. While a large focus has been placed on understanding the microstructural properties from a material science perspective, the vast quantities of data generated during the build have largely been unexploited. This work describes a multi-pronged approach for data discovery, engaging multiple analytic tools as well as a framework to ingest and house the data itself in an effort to identify areas for process improvement and promote the potential for advanced defect detection.

---

95. Selecting the Selector: Using Statistics to Discover New Materials
James Theiler, Los Alamos National Laboratory

We treat the materials discovery problem from a design-of-experiments perspective. We posit a finite set of candidate material compounds, only a few of which have a measured property of interest y (eg, elastic modulus), but all of which have known features x (eg, orbital radii of the individual elements). Although the y values can be approximately predicted from the x values, they cannot be exactly predicted, and we do not know the prediction function a priori. Individual y values can be accurately determined, but the process (which is an experimental measurement) is expensive. The goal is to identify a material from the candidate set with a large value of y, using a small number of experimental measurements.

We consider an iterative approach that alternates between two main steps. One step is to use the available x and y values to estimate a /regressor/ f(x) that predicts a distribution for y; that is, an expected value and an estimated error. The second step uses f(x), applied to all the materials in the candidate set, to /select/ a material for experimental measurement. With a new (x,y) pair, we can go back to the first step, and iterate.

The emphasis of this study is on strategies for selecting the candidate material for the next measurement. This can be cast as a trade-off between selecting the material with largest expected y (exploitation) and selecting the material whose (x,y) values are expected to most improve our model (exploration).

Joint work with Beate Zimmer (Texas A&M University, Corpus Christi), Prasanna V. Balachandran (LANL, T-4), John Hogden (LANL, CCS-3), Kipton Barros (LANL, T-1), Turab Lookman (LANL, T-4)

96. Trajectory Prediction Using Feature Vectors
Andrew Wilson, Sandia National Laboratories

We demonstrate an approach to predicting the destination of a flight in progress using feature vectors. We first construct a database of feature vectors that associate partial trajectories of historical flights with their destinations. Next, we compute a feature vector for a newly observed partial flight. We search for similar feature vectors in the historical database and return the destinations of nearby feature vectors as possibilities. This approach is very fast (logarithmic in the number of historical flights used for comparison) and allows for assignment of probability to each possible destination depending on how frequently it occurs in nearby feature vectors.

Joint work with Benjamin Newton (University of North Carolina at Chapel Hill), Mark D. Rintoul (Sandia National Laboratories), Christopher Valicka (Sandia National Laboratories), Michael Wolf (Sandia National Laboratories)

97. Autoencoder Based Anomaly Detection for Streaming Analytics
Sean Robinson, Pacific Northwest National Laboratory

Some of the most pressing analytics challenges such as cyber security and object recognition often have no-ground-truth and limited-prior-domain-knowledge. A primary goal is to determine when data behavior deviates from the norm, but this is hard to determine with current tools. One way to approach problems without an obvious set of underlying dynamics is to characterize the data using an autoencoder and then examine the learned feature space. We demonstrate the use of an autoencoder to both learn a feature space and then identify anomalous portions without the aid of labeled data or domain knowledge. We use a trained autoencoder to reconstruct incoming data, and then construct a similarity metric between input and reconstruction to determine the data subsets that most deviate from the variations expected from past observations. In this way, we can find unusual sequences and subsets in arbitrary data sets. We have demonstrated this on the standard MNIST handwriting data sets, as well as ElectroCardioGram (ECG) data, to demonstrate the detection of anomalous features without domain knowledge or tagged examples. This method is being expanded to operate on streaming data, to facilitate generation of a real-time streaming indicator of anomalous behavior. We are investigating the use of this technique for a variety of other data challenges including cyber security analysis and real time medical data analysis.

Joint work with Nicole Nichols, Pacific Northwest National Laboratory, Rob Jasper, Pacific Northwest National Laboratory

98. Exploring Simulator Structural Uncertainties Using Quasi-Intrusive Reduced Models
Nathan M. Urban, Los Alamos National Laboratory

Numerical simulators often contain large uncertainties due to modeling choices that cannot be easily expressed in parametric form, such as differences in numerical schemes, physics simplifications, etc. Typically these 'structural' choices are embodied only in the source code that implements them, precluding any kind of automated sampling over the space of possible simulator structures: to explore a different model structure, one must write a new simulation code. We exploit the fact that simulators based on discretized partial differential equations have a common mathematical form: their dynamics can be expressed in terms of a local operator that propagates the state of a grid cell forward in time (e.g. a numerical stencil operator); the dimensionality of this operator is independent of the size of the computational domain or mesh resolution. By inferring this local operator through statistical sampling of simulator output, we can non-intrusively identify the simulator's structure, and perturb this function approximation to automatically generate new candidate structures. To approximate the output of a structurally perturbed model we apply Galerkin projection, an intrusive method of projecting a system's governing equations onto a low-dimensional basis to form an efficient reduced model. When combined with our local system identification approach, this projection method becomes non-intrusive. Our method thereby explores the behavior of different simulator structures in a manner that is non-intrusive, automated, and computationally efficient.

Joint work with Terry Haut (Los Alamos National Laboratory)

99. Modeling of Severe and Fuzzy Uncertainties in Contaminated Groundwater Management

Xiaodong Zhang, Los Alamos National Laboratory

Decision analyses for management of contaminated groundwater sites are facing many challenges. Among them, quantifying uncertainties associated with numerous model parameters, factors and processes is very important since the uncertainties may significantly affect decisions related to groundwater remediation and risk management. Although probabilistic approaches are effective in quantifying random uncertainties, they may encounter difficulties due to lack of enough data for eliciting probability density functions, subjective and vague data/information, and/or severe (deep) uncertainties. Using purely probabilistic methods in such situations may lead to bias or ineffectiveness in the decision making. In this study an integrated modeling approach is proposed to assess and quantify severe and subjective uncertainties associated with groundwater contamination using the information-gap and fuzzy sets theories. A representative case study is developed to demonstrate its applicability for generating scientifically-defensible decision alternatives for groundwater management.