1. An Approach to Characterizing Spatial Aspects of Image System Blur
[P] Jesse Adams, Nevada National Security Site

In imaging applications, such as X-ray radiography, the experimental system setup and data capture process introduce system blur and distortion. In images taken with thick scintillators, a spatially-varying system blur is observed, requiring a non-standard approach to blur and distortion estimation and removal. We demonstrate an approach to estimation of local point spread functions using checkerboard radiographs collected at the Naval Research Laboratory and extend that into a map defining the kernel over the camera field of view.

Joint work with Kevin Joyce, Eric Machorro, Eric Clarkson, Ryan Coffee, Malena Español, Michael Brennan, Jessica Pillow, Daniel Champion, Sean Breckling, Arnulfo Gonzalez, Marylesa Howard. This work was done by Mission Support and Test Services, LLC, under Contract No. DE-NA0003624 with the U.S. Department of Energy and supported by the Site-Directed Research and Development Program. DOE/NV/03624--0683

2. Deployable Predictive Maintenance Strategy Based on Models Developed to Monitor Assets in Nuclear Power Plants
[P] Koushik Araseethota Manjunatha, Idaho National Laboratory

To achieve high capacity factors, the nuclear fleet has over the years relied on labor-intensive and time-consuming preventive maintenance programs to operate and maintain plant systems. This has led to high operation and maintenance costs because it involves manually performed inspection, calibration, testing, and maintenance of plant assets at periodic frequencies, along with time-based replacement of assets, irrespective of condition. This has resulted in a costly, labor-centric business model.

Fortunately, there are technologies that can eliminate preventive maintenance activities by deploying risk-informed predictive maintenance, enabling the transition to a technology-centric business model. The technology-centric business model will drive down costs since labor is a rising cost and technology is a declining cost. This transition will also enable nuclear plants to maintain high -- and possibly even achieve higher -- capacity factors, while still significantly reducing operation and maintenance costs.

This poster presents the research and development performed to enable deployment of a risk-informed predictive maintenance strategy at a nuclear power plant site, based on models developed to monitor the motor-driven pump system as a representative plant asset. This strategy will be adopted by commercial nuclear power plants to replace time-based preventive maintenance strategies, eliminating unnecessary operation and maintenance costs. The poster exhibits the installation of wireless vibration sensor nodes on the circulating water system to enable online vibration monitoring to support development of diagnostic and prognostic models. A framework to scale the presented predictive maintenance strategy across different plant assets at the plant site and across the nuclear fleet is also discussed.

Joint work with Vivek Agarwal, James A. Smith, Vaibhav Yadav, Francis Lukaczyk, Michael Archer, and Palas Harry. The work presented in the poster is being developed as part of a collaborative research effort between Idaho National Laboratory, PKMJ Technical Services, and Public Service Enterprise Group Nuclear, LLC.

3. A Data-Driven Approach to Quantify Major Impacts in Career Decisions
[S] Payton Bartz, Brookhaven National Laboratory

There is a plethora of factors that can contribute to career decisions, such as financial stress, parental influence, and education or job availability. With the introduction of many personal factors, it is difficult to pinpoint a common influence, or influences, which directly affect career choices on the grandest scale. In an effort to understand the most prevalent and holistic influences, a survey was developed to analyze the major influencers in individual career choices and when these influencers carry the most impact. This survey was administered electronically to over 100 participants aged 18 and older, encompassing many different careers and geographic regions. The findings indicate that teachers have the strongest impact on the individual's career choice and that this impact trends upward from primary to secondary education yet plateaus at the tertiary education level. Additionally, these results indicated the most impressionable stages for an individual regarding their academic ambitions. The insights gained from the overall survey results can be used for further research into other contributing factors for career choice, such as why teacher influence peaks at the high school level.

Joint work with Katherine Bachner, Brookhaven National Laboratory.

4. Diagram Image Retrieval using Sketch-Based Deep Learning and Transfer Learning
[S] Manish Bhattarai, Los Alamos National Laboratory

Image retrieval tasks performed on diagram imagery remains a difficult problem in the field of computer vision. As yet, algorithms capable of comparing imagery featuring simple line work like that found in diagram drawings remain unsolved. While deep learning methods continue to excel in the fields of object detection and image classification applied to natural imagery, the application of such methodologies applied to binary imagery such as scientific diagrams/patent images remains unexplored. This work presents a deep learning-based methodology to address this gap. A significant challenge to diagram based imagery is the lack of training data needed. We have devised a method for a visual image search for binary patent images that takes advantage of existing large natural image repositories. Here, we implement the concept of domain generalization by pre-training an unsupervised deep learning model on a large set of natural image generated sketches. Then using transfer learning/domain generalization, we transfer the acquired knowledge to be used on the diagram dataset for latent representation. We also show that even under zero-shot and one-shot conditions, this framework surpasses all the classical retrieval frameworks for binary datasets. We achieved retrieval precision and recall of more than 80% in our experiment. Our experiment results show the effectiveness of deep learning with transfer learning for detecting plagiarism in patent images and querying similar images based on content.

Joint work with Manish Bhattarai, Diane Oyen, Liping Yang, Brendt Wohlberg.

5. Extracting nutrient - ocean state relationships from oceanic biogeochemistry simulations for macroalgae mariculture
[P] Zhendong Cao, Los Alamos National Laboratory

Macroalgae mariculture has the potential to mitigate harmful algal blooms in nutrient rich coastal waters and provide a sustainable biofuel feedstock, as is being evaluated within the Advanced Research Projects Agency–Energy Macroalgae Research Inspiring Novel Energy Resources (MARINER) program. The evaluation of viable mariculture ¦eld sites requires a detailed understanding of nutrients and ocean conditions on site. The fundamental challenge is that existing observational and reanalysis data products do not have nutrient information for these purposes. Nutrients computed within the US Department of Energy's Energy Exascale Earth System Model (E3SM) are at global to regional scale, but mariculture applications require nutrient information at ¦eld scale. To bridge this gap, we build a nutrient prediction model based on calibrated ocean biogeochemstry output from E3SM using a Random Forest Regression (RFR) algorithm. The E3SM output includes the ocean state variables of surface current velocities, temperature, salinity and biogeochemical variables like nutrients at a three hour frequency. The RFR algorithm is trained by daily data to estimate nitrate for the rest of the variables. The optimized RFR model, ¦ne-tuned by grid search technique, shows great performance in predicting the spatio-temporal nitrate distribution, with an Out-of-Bag score of 0.97 in model calibration and Goodness-of-Fitting of R2 = 0.96 in model veri¦cation. These results highlight key factors responsible for nutrient spatio-temporal variability and will be used to approximate nutrient concentrations for remote sensing data and in higher resolution reanalysis products, e.g., HYCOM ocean model forecasts.

Joint work with Phillip Wolfram (LANL), Mathew E Maltrud (LANL), Riley Xavier Brady (University of Colorado Boulder).

6. The role of seasonality and climatic factors in the spatiotemporal transmission dynamics of dengue in Brazil
[P] Lauren A Castro, Information Systems and Modeling, Los Alamos National Laboratory

Although annual DENV outbreaks are observed in all 27 of Brazil's states, outbreak timing can vary by year and region due to the heterogeneity in environmental conditions. Here, we aim to characterize the spatiotemporal spread of DENV and the seasonality of DENV epidemiology in Brazil. We analyzed DENV timing using state-level weekly DENV case counts collected from the Brazilian Ministry of Health for 2010-2016. Using wavelet time series decomposition, we quantified differences in annual outbreak timing between pairs of states and used nonparametric spatial statistics on 88 climatic, remote sensing, and demographic variables to identify factors that may modulate DENV's spatiotemporal dynamics. In addition, we quantified the intensity of DENV seasonality from each state's wavelet spectra and determined factors associated the strength of the seasonal signal using linear regression. Across Brazil, the average difference in DENV outbreak timing spans 30 weeks and follows a counterclockwise pattern, with annual disease activity starting in the Amazon region and ending in the northeastern coastal states. Mantel tests demonstrated that state pairwise correlations in the changes in water content (NDWI) were most strongly associated with coherence in epidemic timing. The intensity of DENV seasonality is strongest in the populous southeastern and northeastern regions of Brazil, and weakest in the Amazon region. Finally, the intensity of DENV seasonality in a state is best predicted by the seasonal intensities of the daily range in temperature, NDWI, the normalized burn ratio, and the percentage of cloudy pixels. This study reveals areas of Brazil where the strength of seasonality and timing of DENV outbreaks is associated with measurable environmental factors. Overall, our results highlight the importance of understanding the complex mechanisms that drive DENV transmission dynamics across diverse geographic regions, and how this knowledge can aid in designing mitigation strategies and informing forecasting efforts.

Joint work with Kaitlyn Martinez(1), Amanda Ziemann (2), Geoffrey Fairchild (1), Carrie Manore(1), Sara Del Valle (1) 1: Information Systems and Modeling, Los Alamos National Laboratory 2: Space Data Science & Systems, Los Alamos National Laboratory.

7. Natural Language Processing for Materials Discovery and Design
[S] John Dagdelen, Lawrence Berkeley National Laboratory

The majority of all materials data is currently scattered across the text, tables, and figures of millions of scientific publications. This poster presents the work of our team at Lawrence Berkeley National Laboratory on the use of natural language processing (NLP) to extract and discover materials knowledge through textual analysis of the abstracts of several million journal articles. With this data we are exploring new avenues for materials discovery and design such as how functional materials like thermoelectrics can be identified by using only unsupervised word embeddings for materials. To date, we have used advanced techniques for named entity recognition to extract more than 100 million mentions of materials, structures, properties, applications, synthesis methods, and characterization techniques from our database of over 3 million materials science abstracts. This poster will also present some of the details on how we are making all of this data freely available to the materials research community through our public-facing website (matscholar.com) and our open-access API.

8. Using Uncertainty To Interpret Machine Learning Predictions
[P] Michael C. Darling, Sandia National Laboratories

Traditionally, machine learning models are assessed using methods that estimate an average performance against samples drawn from a particular distribution. Examples include the use of cross-validation or hold-out to estimate classification error, F-score, precision, and recall. While these measures provide valuable information, they do not tell us a model's certainty relative to particular areas of the input space. Typically there are regions where the model can differentiate between the classes with certainty, and regions where the model is much less certain about its predictions. We demonstrate an approach for quantifying uncertainty in the individual predictions made by supervised machine learning models. We develop an uncertainty measure we call minimum prediction deviation which can be used to assess the quality of the individual predictions made by supervised two-class classifiers. We show how minimum prediction deviation can be used to differentiate between the samples that a model predicts credibly, and the samples for which further analysis is required.

9. Forecasting mosquito populations in the Greater Toronto Area using multispectral remote sensing imagery
[S] Sydney DeMets, Los Alamos National Laboratory

West Nile-Virus (WNV) and St. Louis Encephalitis (SLE) are two of the most common mosquito-borne diseases in North America. WNV and SLE have seemingly random spatial and temporal outbreaks, making outbreaks difficult to predict. However, recent studies have found that mosquito abundance is strongly correlated with WNV and SLE transmission, providing researchers with a starting point for the development of mosquito-borne disease forecasting systems. Mosquito populations are controlled by a variety of environmental variables, including humidity, temperature, vegetation, and available water habitat for breeding. Current mosquito-population forecasting models heavily weigh precipitation inputs, as they are traditionally seen as the best estimate of available breeding space in a region of interest. Although rainfall data are easy to acquire, precipitation data may not be the best estimate of mosquito habitat, as water does not flow evenly across landscapes. Furthermore, precipitation data generally come at a spatial resolution of 800 meters to 2.5 kilometers, and while this resolution can help predict mosquito abundances on large spatial scales, it inhibits the estimation of mosquito populations in urban areas with high landscape heterogeneity. To overcome these limitations, this research explores the use of multispectral imagery for predicting mosquito populations, specifically in the Greater Toronto Area. Multispectral imagery is an attractive data source for predicting mosquito abundance due to its consistent collection and comparatively high spatial resolution (30 meters for Landsat). We derive a weekly time series of standard spectral indices from multispectral imagery over the Greater Toronto Area from 2004 to 2015. We then explore how spectral indices perform as a predictor for combined Cx. restuans and Cx. pipiens mosquito populations, with the ultimate aim of using multispectral imagery to forecast mosquito-borne diseases in highly urbanized areas.

Joint work with Amanda Ziemann, Carrie Manore.

10. Coarse-Grain Cluster Analysis of Tensors With Application to Climate Biome Identification
[P] Derek DeSantis, Los Alamos National Laboratory

A tensor provides a concise way to codify the interdependence of complex data. Treating a tensor as a d-way array, each entry records the interaction between the different indices. Clustering provides a way to parse the complexity of the data into more readily understandable information. Clustering methods are heavily dependent on the algorithm of choice, as well as the chosen hyperparameters of the algorithm. However, their sensitivity to data scales is largely unknown.

In this work, we apply the discrete wavelet transform to analyze the effects of coarse-graining on clustering tensor data. We are particularly interested in understanding how scale effects clustering of the Earth's climate system. The discrete wavelet transform allows classification of the Earth's climate across a multitude of spatial-temporal scales. The discrete wavelet transform is used to produce an ensemble of classification estimates, as opposed to a single classification. Using information theory, we discover a sub-collection of the ensemble that span the majority of the variance observed, allowing for efficient consensus clustering techniques that can be used to identify climate biomes.

Joint work with Phillip J. Wolfram, Katrina Bennett, Boian Alexandrov - Los Alamos National Labs.

11. Applying Deep Learning on Dark Fiber Network for Near-Surface Characterization
[P] Vincent Dumont, Lawrence Berkeley National Laboratory

Distributed Acoustic Sensing (DAS) is a new technique that utilizes unused fiber-optic cables and repurposes them as seismic sensors (Daley et al., 2013). Using the imperfections of the fiber and the Rayleigh scattering of the light propagating through the cable, geophysicists have been able to use optical fiber lines as a distributed strain sensing network where consecutive segments of the fibers can be probed to monitor seismic activity and perform all sorts of geophysical studies. For instance, DAS technique can be used successfully in doing near-surface characterization (Ajo-Franklin et al., 2019) using a data analysis approach known as ambient noise interferometry. The technique consists of using the strain measurement time series from each segment of the fiber to extract surface wave velocity information and map shallow structural profiles. The standard input data used for this research consist of 3D images where one axis is time, another is the segment index, and the third corresponds to the strain value of a given region in the fiber and at a given time. When heavy loads such as cars and trains are moving around the fibers, this produces tiny earthquakes which can be seen in the form of surface wave signals in the data. The heavier the moving object on the ground is and the more energetic the surface wave signal will be. At the moment, the most efficient way to do ambient noise interferometry is to only use data where highly-energetic surface wave signals produced by moving trains are available. However, this approach is discarding the majority of data present in storage and researchers have been looking for innovative ways to benefit from this big dataset and study it in an automated fashion. In this work, we present a machine learning driven approach that combines both unsupervised and supervised learning where the former is used for clustering of surface wave signals and the latter is used for classification of unknown incoming DAS data. We applied our highly-parallelized algorithm on a large dataset of DAS data to extract surface wave probability information. This can then be used to significantly relax the constraints on the stacking requirements of cross-correlated data needed to obtain shear-wave velocity profiles. By applying machine learning to the entirety of the DAS data, without making any selection, we demonstrate that we can outperform the current analysis approach for ambient noise interferometry.

Joint work with Veronica Rodriguez Tribaldos (LBNL), Benjamin Erichson (UC Berkeley), John K. Wu (LBNL), Jonathan Ajo-Franklin (LBNL, Rice University), Michael Mahoney (UC Berkeley).

12. Precision Aggregated Local Models
[S] Adam M. Edwards, Virginia Tech

Gaussian Process (GP) models have long been used as a flexible method for nonparametric regression. Despite their accuracy, they are infeasible for larger data sets due to the scaling of the computational burden (O(N^3)) and the storage requirements (O(N^2)). Typical methods to combat intractability have focused on splitting the larger GP into smaller problems that can be treated as functionally independent either by partitioning the domain space, or partitioning the data themselves into discrete sets. These methods get around the functional dependence of the individual GP models they use by assigning no weight to most of the models during prediction, or exploiting global models learning theoretically the same surface respectively, however each solution has its drawbacks. Partition models maintain accuracy, but lose continuity at the boundaries. Typical averaged models, on the other hand, maintain absolute continuity in both the mean and variance surface while over-smoothing the function as a whole. Using Local Approximate Gaussian Processes (LAGP) as a method to build local experts, the Precision Aggregated Local Models (PALM) approach bridges the gap between these two approaches to create a locally accurate global model that maintains absolute continuity. We address the functional dependence of individual GP models, creating a unified surface with drastically reduced computational burden O(Nc^2 + Nc * nc^3), and storage requirement O(Nc^2 + Nc*nc^2), where Nc, nc << N.

Joint work with Robert B. Gramacy.

13. Model-Based Optimization and Bayesian Learning from Dynamic Membrane Characterization Experiments
[S] Elvis A. Eugene, University of Notre Dame

Membrane systems are superior to current separations processes because of lower footprints (compactness) and lower mass transfer limitations (rapidity) and are therefore an exciting field of research for sustainable technology. Specifically, membrane systems show great promise in remediating man-made messes via applications in purification of contaminated water and recycling of spent lithium-ion batteries.

Despite this promise, fundamental mechanisms that govern interfacial and transport phenomena for membranes are not well understood, and thereby pose a significant gap in knowledge that inhibit their translation to scalable technologies. In this work, first, we discuss a statistical learning framework for the analysis of dynamic diafiltration experiments. We use timeseries data to characterize membranes by estimating model parameters governing interfacial and transport phenomena across membranes. We argue that analyzing timeseries data from dynamic experiments will allow for faster characterization of membranes and capture non-equilibrium effects encountered in the process. The best fit model is computed by minimizing the weighted sum of squares error between model predictions and experimental data. We then compute the covariance matrix and assemble the joint elliptical confidence region, which captures the uncertainty in the fitted parameters and indicates the quality of fit.

Next, using a simple example of ballistic firing, we demonstrate Bayesian hybrid models that leverage data and domain knowledge to account for model inadequacy which enables the optimal design of experimental conditions. Bayesian hybrid models fuse physics-based insights with machine learning constructs to correct for systematic bias. Systematic bias in models generally arise from simplifying assumptions which do not account for underlying physical phenomena in a process. In an illustrative ballistic firing case study, we use the posterior distribution of the hybrid model to compute optimal firing conditions to hit a target using a single stage stochastic program. We show that hybrid models outperform pure physics or machine learning models in a workflow for data-informed decision-making under parametric and model-form (epistemic) uncertainty. Finally, we compare the frequentist parameter estimation workflow with Bayesian model calibration for dynamic experiments. We argue that the posterior parameter distributions obtained from the Bayesian methods lend themselves directly to decision making under uncertainty using stochastic programs, which can be used for the optimal design of future experiments to maximize user defined goals such as minimizing uncertainty (exploration) or maximizing a cost function (exploitation). We conclude by discussing our immediate next steps in developing our learning and hybrid modeling frameworks to realize our ultimate vision of integrated engineering frameworks that span molecular to systems scales.

Joint work with Xinhong Liu, William A. Phillip, Alexander W. Dowling University of Notre Dame.

14. Understanding Dynamics of Transportation Traffic Flow Around O'Hare International Airport
[P] Arindam Fadikar, Argonne National Laboratory

In this work, we aim to characterize the traffic pattern around the O'Hare airport in the Chicago metropolitan area by combining and analyzing (heterogeneous) data arising from different sources. The data primarily consists of recorded trips from a (unknown) sample of anonymous vehicles providing geocoded location of the vehicles at equally spaced temporal intervals along with instantaneous speed and direction, spanned across a week; and total vehicle counts on road at few predefined locations. We demonstrate our ability to interpolate geocoded trips at finer spatio-temporal resolution by infusing them with underlying road network (aka map-matching). This network view of the data allows one to perform queries on the traffic pattern at any given time, on any specific road with a direction, which is pivotal in the characterization of the changes in the trip distribution statistics and how these depend on features such as roads, time of the day, day in week, and/or weather-related events. By characterizing the trips alone, several interesting inference can be made, e.g. distribution of types of vehicle (such as rideshare, personal, commercial etc.) that corresponds to each recorded trip. Inferred flow distribution from this data later serves as a validation for large-scale traffic simulation for optimization and forecasting.

Joint work with Stefan Wild (ANL), Joshua Auld (ANL).

15. Understanding the Effects of Tapering on Gaussian Process Regression
[S] Juliette Franzman, Lawrence Livermore National Laboratory

Gaussian process regression is a computationally expensive machine learning algorithm that requires the solution of a system of linear equations with a dense covariance matrix. We propose a modification that uses tapering to introduce sparsity into the covariance matrix followed by matrix reordering schemes to reduce the bandwidth of the sparse matrix. We also apply iterative refinement to our approximate solution in order to recover some of the accuracy lost to tapering. Our numerical tests showed that this modification works best with datasets that have many small values in the covariance matrix. This restricts the usefulness of the modification, so other techniques to reduce computational time may be worth considering in the future.

Joint work with Chandrika Kamath Lawrence Livermore National Laboratory.

16. Statistical Forecasting and Stochastic Programming for Energy Market Participation
[S] Xian Gao, University of Notre Dame

The DOE-funded Institute for the Design of Advanced Energy Systems (IDAES) seeks to establish open-source frameworks for data-informed optimization of existing and future energy conversion systems. IDAES offers unprecedented analysis capabilities that span seconds to decades timescale by integrating physics-based dynamic optimization, data reconciliation, conceptual process design, energy infrastructure capacity expansion. As part of the IDAES project, we are developing new methodologies to identify and exploit emerging opportunities day-ahead, real-time and ancillary service electricity markets. Undoubtedly, emerging technologies and greater adoption of renewable sources are driving energy markets to become more dynamic [1]. Through IDEAS, we seek to guide the development of future energy conversion technologies to be best aligned with the needs of future power systems infrastructure in the US.

A standard technique to quantify the profitability of these technologies in the previous studies is to calculate the maximum possible revenue in retrospect, ignoring uncertainty from electricity prices, weather and other sources [2]. In this work, we quantitatively explore how price uncertainty impacts optimal market participation strategies and resulting revenues. Specifically, we benchmark 2-stage stochastic programming formulations for self-schedule and bidding market participation modes in a receding horizon model predictive control framework [3,4]. To generate probabilistic price forecasts, we propose an autoregressive Gaussian process regression model and compare three sampling strategies, i.e. Monte Carlo sampling, uniform sampling and contour sampling [5]. As an illustrative example, we study a price-taker generation company with six unique generation units using historical price data from CAISO (California market). We show that self-schedule is sensitive to the error in the forecast mean and is the best to pair with Monte Carlo sampling which has the lowest root mean square forecasting error, whereas bidding requires price forecasts, such as contour sampling, that cover extreme events (e.g., tails of the distribution). We benchmark realized market revenue against optimal bidding with perfect information and find static bid curve, time-varying bid curve, and self-schedule modes recovery 95.29%, 94.85%, and 84.87% of perfect information revenue, respectively. This result shows a reliable method to quantify the profitability of energy systems and it should be incorporated in the design process of future energy systems.

Joint work with Alexander W. Dowling, University of Notre Dame.

[1] D. J. Chmielewski, 'Smart grid the basics-what? why? who? how?,' Chemical Engineering Progress, vol. 110, no. 8, pp. 28-33, 2014.

[2] Dowling, A. W., Kumar, R., & Zavala, V. M. (2017). A multi-scale optimization framework for electricity market participation. Applied Energy, 190, 147-164.

[3] Ierapetritou, M. G., Wu, D., Vin, J., Sweeney, P., & Chigirinskiy, M. (2002). Cost minimization in an energy-intensive plant using mathematical programming approaches. Industrial & Engineering Chemistry Research, 41(21), 5262-5277.

[4] Plazas, M. A., Conejo, A. J., & Prieto, F. J. (2005). Multimarket optimal bidding for a power producer. IEEE Transactions on Power Systems, 20(4), 2041-2050.

[5] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

17. Multi-spacecraft Observations of Substorm Injections Inside, Outside, and At Geosynchronous Orbit
[S] Sophie Graf, Los Alamos National Laboratory

The concept of a substorm 'injection boundary' was introduced in the 1970s by Mauk, McIlwain and others. The standard picture that remains today is of a spatial boundary that separates a region where dispersionless injection signatures are observed and a region where the injections are dispersed by energy-dependent drift away from the injection region. Multi-energy particle drift tracing confirms that the statistical, single-satellite picture remains valid (but variable) in individual substorm injection events but, important refinements to this picture have been made. They include the fact that the dispersionless injection region is not time stationary but propagates Earthward (at 100s km/s in the magnetotail but only 10s km/s inside geosynchronous orbit); the linking of the injection region to a region where fast Earthward flows break and diverge; the observation of time-dependent dynamics behind (within) the injection region; differences in the electron and ion injection regions; etc.

While the injection boundary paradigm remains robust, current observations offer an unprecedented array of satellites capable of testing and refining the injection boundary picture. Those satellites include, in particular, Van Allen Probes, MMS, THEMIS and the LANL-GEO and LANL-GPS constellations. In this poster we compare simultaneous observations of injections at multiple radial and azimuthal separations. Injections at and inside geosynchronous orbit confirm that the probability of observing an injection increases with radius and that injections inside geosynchronous orbit generally propagate Earthward. We also find that the relationship between injections observed at and outside geosynchronous orbit is considerably more complex. One clear result is that the majority of injections observed outside geosynchronous orbit are observed later than their geosynchronous counterparts.

Joint work with Geoff Reeves (LANL, The New Mexico Consortium), Drew Turner (Aerospace Corporation), Ian Cohen (Johns Hopkins Applied Physics Lab), Christine Gabrielse (University of California Los Angeles), Barry Mauk (Johns Hopkins Applied Physics Lab), Mike Henderson (LANL, The New Mexico Consortium).

18. Classification of Dissolution Events at Radiochemical Processing Facility using Effluents Measurements
[P] Christopher Greulich, Oak Ridge National Laboratory

We address the problem of inferring dissolution events at a nuclear processing facility during a single isotope production campaign using radiation measurements of effluents at an off-gas stack. We utilize datasets collected at the Oak Ridge National Laboratory's Radiochemical Engineering Development Center (REDC), which is a multipurpose radiochemical processing facility involved in the production of a variety of radioisotopes. The abatement system of REDC is instrumented with a high purity germanium (HPGe) detector that records gamma spectra of the effluents. Using the spectra, activity levels of 15 radionuclides, including isotopes krypton, xeon, and iodine, are computed. The time series of these isotopic measurements along with Boolean labeling of plutonium-dissolution events at 1 hour periods are used to design, train, and test a suite of classifiers and fusers.

We utilize a diverse set of classifiers, namely, ensemble of trees, support vector machine, naive Bayes, linear discriminant, and k-nearest neighbor that are based on different design principles which results in varying prediction performance and biases. To mitigate this effect, we utilize classifier fusers or meta learners - such as Chow's rule, meta ensemble, and meta linear discriminant - to combine the outputs of individual classifiers, which generally yielded performance as good as the best performing individual classifier.

Furthermore, to improve the classification performance, we pre-process the isotopic measurements using two transformations prior to training: a convolution operator and a moving average filter. A convolution filter using the Hilbert transform potentially captures the effects of non-stationary structural components that are mechanically or pneumatically coupled to the measurement system, which in turn are reflected in the frequency domain.The moving average filter exploits the nature of a Poisson distributed random processes; appropriately long time windows reduce the uncertainty of measurements, and thus improve the classifier performance, since our classifiers do not explicitly consider the measurement uncertainty. Resulting from an analysis of the respective decay chains and isotopic half-lives, a measurement window of 24 hours is identified.

In general, classifier performance improved with filtering and classifier fusion. Specifically, the Hilbert transform improved the area under the receiver operating characteristic curve (AUROC) for the meta linear discriminant classifier by 9.4% relative to the original AUROC. The best individual classifier, k-nearest neighbor, with the 24 hour average filter resulted in 24.6% improvement relative to the unfiltered case. This result is further improved by the meta linear discriminant classifier which improved relative to the unfiltered case by 26.2% with a final AUROC of 99.18%.

Joint work with Nageswara S. V. Rao, Satyabrata Sen, Jason Hite, Ken Dayman, Andrew Nicholson, Riley Hunley, and Jared Johnson.

19. Emulating and Interpreting Spatial Deposition Simulations
[P] Nipun Gunawardena, Lawrence Livermore National Laboratory

In the event of a hazardous material emission in the atmosphere due to an industrial accident, terrorist attack, or other event, atmospheric researchers often run physics-based atmospheric transport and dispersion models to predict the extent and variation of the contaminant spread. These predictions are subject to uncertainty due to uncertainty in atmospheric models and data. Ensembles of simulations can be used to estimate uncertainty, but running large ensembles is often too time consuming and resource intensive, even using large supercomputers. Here we show that a hybrid linear/logistic regression model can be used to successfully emulate spatial patterns from these physical models with as few as fifty training examples. The use of logistic regression allows for probabilistic predictions which are easily understood. The use of linear regression opens up avenues of exploration regarding interpretability -- the presented model can be used to find which physics schemes are most important and which spatial areas are most problematic. Ultimately, this work will allow atmospheric researchers to produce potential contamination scenarios with uncertainty estimates faster than previously possible, aiding public servants and first responders.

Joint work with Giuliana Pallotta and Donald D. Lucas. This work is released under number LLNL-ABS-801735 and was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC.

20. Modular Bayesian calibration of plasma compositions
[S] Claire-Alice Hébert, Stanford University

The Mars Rover carries onboard an instrument, ChemCam, designed to measure the composition of surface rocks using laser-induced breakdown spectroscopy (LIBS). The spectra from this instrument must be analyzed to identify and quantify the component elements in the source sample, as well as their proportions. This process is called disaggregation, and is complicated by so-called matrix effects which influence the relative height of emission lines in nonlinear ways as a result of combinations of elements. The time intensive plasma physics code ATOMIC, developed by James Colgan at LANL, has been used to model these spectra, but using it for forward models is intractable given the large parameter space to explore. Emulators have been proposed as a faster solution for this analysis. We discuss a proof of concept emulator for disaggregating two element sodium copper compounds, built with Gaussian processes after using principal components analysis (PCA) to reduce the dimensionality of the spectra. The training and testing datasets were simulated with ATOMIC using a latin hypercube design. After testing the performance of the emulator for reconstructing test spectra, we perform modular Bayesian calibration to recover the composition of several test spectra.

Joint work with Earl Lawrence (LANL Statistical Sciences CCS-6), Kary Myers (LANL Statistical Sciences CCS-6), James Colgan (LANL theoretical devision T-1).

21. Quantum vs Neuromorphic: Solutions to Post Moore's Law Computing
[S] Kyle Henke, Los Alamos National Laboratory

Throughout the scientific community there is growing concern of humanities increasingly rapid approach towards the theoretical limits of classical computation, better known as the end of Moore's law. In-spit of this knowledge, reliance on machine learning and autonomous products has exponentially grown in the age of Big Data, and hence, the exploration of novel computational platforms which can overcome the current deficiencies, must be embarked upon. Because the human mind is incredibly energy efficient, and quantum mechanics provides the theoretical foundations for extremely fast solutions to difficult optimization problems, a comparison of neuromorphic and quantum annealing techniques is the first and most promising step towards satisfying the growing thirst for computing abilities into the future.

Classical computing can be divided into two categories, depending on whether the underlying circuits are digital or analog. Digital logic gates are universal, allowing the construction of computing architectures capable of running any valid program. Analog computers, conversely, use the dynamical evolution of a physical system to perform a given computation. While analog computers can be extremely fast and power efficient, the noise associated with the evolution of such systems causes difficulty when programming and often represents a major limitation.

Neuromorphic processors, drawing inspiration from biological brains, comprise a class of ultra low-power analog devices that are capable of self-organizing in response structured input. In this sense, neuromorphic processors are able to program themselves, potentially alleviating the major limitation of analog computing devices. As we require computers to exhibit greater autonomy and intelligence, and the focus of computing applications shifts toward machine learning and machine intelligence, analog neuromorphic processors are likely to play an increasingly prominent role.

The question of analog vs digital transfers over into quantum computing regime as well. Like their classical counterparts, quantum logic gates can, in principle, enable the construction of quantum computers capable of running any valid program. Also like their classical counterparts, quantum analog computers, such as quantum annealing machines, are physical systems in which the dynamical evolution of the system performs the desired computation. Both analog and gate-based approaches to quantum computing seek to exploit quantum entanglement, superposition, and other quantum effects to solve problems that would otherwise be intractable using a purely classical approach, a goal known as quantum supremacy.

In this work, we compare the abilities of emerging hardware (D-Wave quantum annealing and Intel Loihi spiking neuromorphic) with that of contemporary microprocessors. The sparse coding problem is solved for dimensionally reduced benchmark fashion MNIST data set by employing the neurally inspire LCA.

Joint work with Ben Migliori Garrett Kenyon.

22. AdaOja: Adaptive Learning Rates for Streaming PCA
[S] Amelia Henriksen, Oden Institute, University of Texas at Austin

Oja's algorithm has been the cornerstone of streaming methods in Principal Component Analysis (PCA) since it was first proposed in 1982. However, Oja's algorithm does not have a standardized choice of learning rate (step size) that both performs well in practice and truly conforms to the online streaming setting. We propose a new learning rate scheme for Oja's method called AdaOja. This new algorithm requires only a single pass over the data and does not depend on knowing properties of the data set a priori. AdaOja is a novel variation of the Adagrad algorithm to Oja's algorithm in the single eigenvector case and extended to the multiple eigenvector case. We demonstrate for dense synthetic data, sparse real-world data and dense real-world data that AdaOja outperforms common learning rate choices for Oja's method. We also show that AdaOja performs comparably to state-of-the-art algorithms in the same streaming PCA setting. This work improves an important foundation not only for streaming PCA but for its variants--including streaming kernel PCA.

Joint work with Rachel Ward, Oden Institute, University of Texas at Austin.

https://github.com/aamcbee/AdaOja https://arxiv.org/abs/1905.12115

23. Parla: Heterogeneous Codes in Python
[S] Ian Henriksen, University of Texas at Austin

Parla is a new Python-based tasking system focused on orchestrating coarse-grained kernel calls on heterogeneous systems. It relies on ongoing work in the scientific Python ecosystem from projects like NumPy, CuPy, and Numba to provide a high-level programming environment for computational science that makes it easy for users to leverage diverse computing platforms. Parla enables people to write efficient heterogeneous codes entirely in Python while also providing easy interoperability with existing high-performance libraries and simulation pipelines. The Parla orchestration layer is built around a resource aware dynamic scheduler that avoids oversubscribing compute resources and running out of memory. Work toward enabling automatic data movement between devices, managing the computational resources used by external libraries, and providing a unified way of expressing sparse and semi-structured data is ongoing.

24. Microstructural analysis of biofuel wood pellets using X-ray computed micro-tomographic imaging
[S] Julia Hestenes, Advanced Light Source, Lawrence Berkeley National Laboratory

This project presents the development of an image processing procedure for synchrotron X-ray computed micro-tomographic (micro-CT) image scans of biofuel wood pellets as well as analyses of the results with respect to the pellet manufacturing process. Micro-tomographic imaging of these pellets reveals unique crack structures and structural asymmetry along the axis of pelletization. Methods of extracting information of interest from micro-CT 3D image sets are intricate and unique to the object being studied. This work discusses image analysis techniques used to segment the images and ultimately characterize the microstructural features and geometry of several biofuel wood pellet specimens. Novel findings are presented which include the porosity and two-dimensional maps of the cracks of the pellets, showing the asymmetry and variability of pellet microstructure.

Joint work with Natalie Larson and Dilworth Parkinson, Beamline 8.3.2 at Lawrence Berkeley National Laboratory. Mikael Thyrel and Torbjorn Lestander at the Swedish University of Agricultural Sciences.

25. Methods for determination of dimensionality of latent features: Principal Angles and PCAk
[S] Miguel A Hombrados, Los Alamos National Laboratory

Principal components analysis (PCA) and Independent component analysis (ICA) are well-known methods in machine learning for finding a representation of lower dimensionality of the data in a latentspace. However, none of them allows to determine what is the effective or intrinsic dimensionality of that latent space. In this paper, we propose a set of methods based on resampling techniques, that allows us to find the effective dimensionality in an automatic way, for both PCA and ICA. We perform an extensive comparison between our methods and other prominent techniques for automatic determination of the latent dimensionality with a probabilistic formulation, showing how our methods are able to reach comparable performances without the assumption of any prior distribution over the latent components. Finally, we make an analysis about the parallelisms existing between our methods and other other probabilistic methods in light of the experimental results.

Joint work with Vangara, Raviteja,Skau, Erik West, Alexandrov, Boian.

26. Damage Detection in Metallic Additively Manufactured Parts using In-Situ Steady-State Ultrasonic Response Data

[S] Erica M. Jacobson, Los Alamos National Laboratory

The increasing implementation of additively manufactured (AM) parts into safety critical applications within a wide range of industries is accelerating the demand for non-destructive evaluation (NDE) as a means of quality control (QC) for defect detection during the build process. An adaptation of acoustic wavenumber spectroscopy - an ultrafast NDE technique that has successfully been utilized in a wide range of applications, from weld cracks to delaminations in carbon-fiber reinforced panels - was implemented into a direct laser metal sintering (DLMS) machine for a 304L stainless steel build. A scanning laser Doppler vibrometer collected the direct-part response between each print layer, resulting in a unique, three-dimensional inspection volume of the entire build. This work describes the data analysis techniques used to separate each built component from the surrounding unsintered powder, as well as the statistical measures and machine learning concepts used to identify features found in the response data to locate regions of damage.

Joint work with Peter H. Fickenwirth, Eric B. Flynn, & Adam J. Wachtor (Los Alamos National Laboratory).

27. An Integration of Imaging & Genetic Data for Deep Learning of Brain Diseases

[S] Kristen Nalani Knight, University of Georgia

The goal of the emerging field of imaging genetics is to combine the neuroimaging power of fMRI and the sequencing of the complex human genome to produce a unified approach to understanding the acquisition and progression of psychiatric illnesses. Despite the work towards discovery in the past decade, the groundwork of statistical methodology for imaging genetics remains in its infancy. Numerous challenges exist for this BIG data problem, such as easing the computational burden, minimizing false positives, all while accounting for human heterogeneity. I will outline the current direction of imaging genetics research and highlight the short-comings of past studies through simulation and replication studies. Then, I will introduce my own approach by providing a applicable framework for imaging genetics analyzes. Emphasis will be placed on dimension reduction and candidate-gene selection for genome-wide association studies with a neuroimaging phenotype via machine learning approaches. Data gleaned from the Alzheimers Disease Neuroimaging Initiative (ADNI) are the motivation for my research and will be analyzed. Results will be compared among healthy, mild cognitive impaired and Alzheimers Disease diagnosed subjects.

Joint work with Liang Liu, Nicole Lazar.

28. From Fourier to Koopman: Spectral methods for long-term forecasting of high dimensional systems

[P] Henning Lange, University of Washington

We propose spectral methods for long-term forecasting of signals stemming from linear as well as non-linear dynamical systems that preserve energy. For linear signals, we introduce an algorithm with similarities to the Fourier transform that however does not rely on periodicity assumptions and therefore allows for forecasting. We then extend this algorithms to handle non-linearities by making use of Koopman theory. The resulting algorithm performs a spectral decomposition in a non-linear and data-dependent basis. The optimization objective for both algorithms is highly non-convex, however expressing the objective in frequency domain allows us to compute global optima of the error surface in a scalable and efficient manner, partially by exploiting the computational properties of the Fast Fourier Transform. The efficacy of the algorithms is evaluated on synthetic experiments as well as in the context of predicting signals in the realms of power systems and fluid flows.

Joint work with Steven L. Brunton, J. Nathan Kutz.

29. Spatio-temporal modelling of the FNET/GridEye data

[P] Amanda Lenzi, Argonne National Laboratory

Understanding power system dynamics is essential for inter-area oscillation analysis and detection of grid instabilities. The FNET/GridEye GPS-synchronized wide-area frequency measurement network provides an accurate picture of the overall real-time operational condition of the power system dynamics. We propose to model measurement-based frequency disturbance recorders (FDR) data from the US Eastern Interconnection (EI) with a spatio-temporal statistical model. We predict the frequency data not only where measurements are available but also at locations without observations. This type of scenario is relevant to problems where measurement data are inaccessible, whereas decision support applications require an overview of the power grid at all locations over a region. The spatio-temporal correlations in our model are captured by a Gaussian field that evolves in space with a Matérn covariance and in time with an autoregressive dynamical dependence. Based only on data from a few locations, we construct maps of frequency for the entire EI at present and future times. The preliminary results indicate that the proposed method is useful in estimating power system dynamic response from limited phasor measurements and holds promise for the prediction of instability caused by cascading outages.

30. Large Environmental and Demographic Data Sets in Models for Mosquito Borne Disease Risk in Brazil
[S] Kaitlyn Martinez, Colorado School of Mines

The spread of mosquito-borne diseases is complex and direct measurements of the fundamental mechanisms of spread is an onerous task. Therefore, we must turn to proxy data to gain insights into these systems. Fortunately, fewer limitations in data collection and storage have resulted in an abundance of rich, diverse, and dynamic data sources that measure various aspects of mosquito-borne disease spread. Weather station measurements and remote sensing of vegetation health can elucidate the status of the mosquito habitat, while demographic indicators can provide information about the impact of the manmade infrastructure on mosquito spread. Despite the abundance of proxy data, there are many challenges due to the heterogeneity in these data streams. For example, it is difficult to determine which variables, or collection of variables, are useful given a large number of observations and potential redundancy. In order to address these challenges, we developed an iterative dimension reduction method using hierarchical clustering to decrease the number of variables, while maintaining the intrinsic information of the full data set. Our method is applied to dengue transmission in Brazil and Ecuador and is tailored to maintain biological interpretation of variable sets driving mosquito-borne diseases such as demographic and environmental data. We find that socioeconomic factors, temperature, and levels of healthy vegetation are highly predictive of dengue incidence. Our results are consistent with previous studies that have shown that each of these factors is impacted by climate change in ways that will further increase mosquito-borne disease incidence around the world. Our study results can inform short-term prevention strategies as well as long term public health campaigns focused on reducing the overall burden of mosquito-borne diseases.

Joint work with Los Alamos National Lab: Sara Del Valle, Geoffrey Fairchild, Amanda Ziemann, Nidhi Parikh, Carrie Manore University of Notre Dame: Amir Said Siraj.

31. Improving Multi-Modal Analytics: Building Interfaces Using R
[S] Tanvi Mehta, Los Alamos National Laboratory

The overall goal of this project is to develop effective analytical methods that fuse data from multiple modalities to describe the patterns of life at a nuclear facility. Data is collected from many disparate modalities, including seismic, acoustic, radiation and EM, making it difficult to utilize. To increase utilization, we created functionality in R to use metadata to directly query data records stored in Echo, an internally developed data management tool. We also developed R functionality to filter data from these Echo records, generate plots, and create a Cinema viewer, an interactive image-based tool that allows the user to easily visualize desired subsets. Together, these R interfaces will allow for improved analysis of the targeted non-proliferation data.

Joint work with Jenna Korobova (Los Alamos National Laboratory).

32. Fast Increased Fidelity Approximate Gibbs samplers for Bayesian Gaussian process regression
[S] Kelly R. Moran, Duke University

Gaussian processes (GPs) are common components in Bayesian non-parametric models. Their use is supported by efficient sampling algorithms, a rich methodological literature, and strong theoretical grounding. However, due to their prohibitive computation and storage demands, the use of exact GPs in Bayesian models is limited to problems containing at most several thousand observations. Computational and storage bottlenecks arise when sampling the GP. Sampling requires a matrix inversion and the Cholesky factorization of the conditional covariance matrix; these operations scale cubically with n, the number of unique inputs. Storage of individual matrices scales quadratically, and can quickly overwhelm the resources of most modern computers for larger problems. To overcome these bottlenecks, we develop a sampling algorithm using H-matrix approximation of the matrices comprising the GP posterior covariance. These matrices can approximate the true conditional covariance matrix within machine precision and allow for sampling algorithms that scale at O(n log^2 n) time and storage demands scaling at O(n log n). We also describe how these algorithms can be used as building blocks to model higher dimensional surfaces at O(d n log^2 n), where d is the dimension of the surface under consideration, using tensor products of one-dimensional GPs. Though various scalable processes have been proposed for approximating Bayesian GP inference when n is large, to our knowledge, none of these methods show that the approximation's Kullback-Leibler divergence to the true posterior can be made arbitrarily small and may be no worse than the approximation provided by finite computer arithmetic. We describe H-matrices, give an efficient Gibbs sampler using these matrices for one-dimensional GPs, offer a proposed extension to higher dimensional surfaces, and investigate the performance of this fast increased fidelity approximate GP, FIFA-GP, using both simulated and real data sets.

33. Error Propagation in Recursive Multi-fidelity Emulators
[P] Amanda Muyskens, Lawrence Livermore National Laboratory

In order to estimate physics parameters using Bayesian calibration, it is often necessary to build an emulator to approximate the output of expensive physics simulations for new parameter combinations. In many cases, the physics simulations used to train the emulator can be run at multiple grid resolutions, where dense grid simulations are slower but more realistic. In an effort to reduce the computational cost, a combination of simulations with different resolutions can be combined to develop a more accurate emulator within a fixed training time. One popular such model is the recursive co-kriging multi-fidelity model, where each level is represented autoregressively as the output of the closest grid resolution with the addition of an independent Gaussian process to model the level discrepancy (Kennedy and O'Hagan, 2000). The levels of fidelity in which one takes observations is assumed to be pre-determined, but this is often not the case where practically any grid resolution can be selected. However, prediction error in each layer of fidelity produces more error in the overall multi-fidelity model so it may not be ideal to incorporate information from repetitive layers or from layers that are too different from the high-fidelity surface. Therefore, we describe a generalization of this multi-fidelity model, demonstrate how prediction error in each level of fidelity effects error in the emulator predictions, and describe a validation scheme for the configuration of the model after training data is collected.

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

Joint work with Kathleen Schmidt, Lawrence Livermore National Laboratory. LLNL-ABS-800957

Kennedy, Marc C., and Anthony O'Hagan. 'Predicting the output from a complex computer code when fast approximations are available.' Biometrika 87.1 (2000): 1-13.

34. Comparison of Emulation Techniques for Bayesian Calibration
[S] Samuel Myren, Los Alamos National Laboratory

Computer simulation, or modeling, of physical phenomena is a vital task for problems encountered in the Department of Energy due to ethical, monetary, and physical constraints associated with obtaining experimental data. To ensure alignment between computer simulation and experimental or natural phenomena, the simulation must be calibrated by adjusting simulation input parameters which entails running the simulation thousands of times at various parameter settings. Since simulations may take hours to weeks to complete a single run and the number of input parameters may be large, calibration through direct sampling of the simulation becomes computationally infeasible. Therefore, an approximation to the simulation, known as an emulator, is used during calibration instead of the simulation to allow for faster sampling at the expense of some error. This research focuses on the emulator's influence in calibration. Three methods of emulation: deep neural networks, Bayesian Additive Regression Trees, and Gaussian processes, are applied through Bayesian calibration as described by Kennedy and O'Hagan (2002) to three different datasets. The datasets are representative of the type of work performed at Los Alamos National Laboratory where experimental and simulation sample sizes are relatively small. These simulations include an aluminum flyer plate model (n = 1000), properties of atomic nuclei (n = 200), and astronomical data (n = 30). For each emulator, the resulting posterior distribution of the model parameters, the computational efficiency, and the predictive capabilities are compared.

Joint work with Earl Lawrence, Los Alamos National Laboratory.

35. On function approximations using Fourier neural networks
[P] Marieme Ngom, Argonne National Laboratory

Recently, machine learning tools in particular neural networks have been widely used to solve differential equations. One main advantage of using machine learning, in this case, is that one does not need to mesh the computational domain and can instead randomly draw data points to solve the differential equations of interest. In this work, we propose a simple neural network to approximate low-frequency periodic functions or seek such solutions of differential equations. To this end, we build a Fourier Neural Network (FNN) represented as a shallow neural network (i.e with one hidden layer) based on the Fourier Decomposition. As opposed to traditional neural networks, which feature activation functions such as the sigmoid, logistic, ReLU, hyperbolic tangent and softmax functions, Fourier Neural Networks are composed using sinusoidal activation functions. We propose a strategy to initialize the weights of this FNN and showcase its performance against traditional networks for function approximations and differential equations solutions.

36. Modeling and approximate inference in complex systems with hybrid dynamic Bayesian networks
[P] Phan Nguyen, Lawrence Livermore National Laboratory

Analyzing and modeling high-dimensional temporal data streams remains of great interest to many scientific and engineering applications. In data analytics problems such as autonomous vehicle control, target tracking, biological sequence analysis, signal processing, and fault diagnosis, multiple sensors may provide sequences of heterogeneous information with which one may use to make predictions about a system and its underlying complex dynamics. These tasks require models and algorithms that can efficiently synthesize information from multiple dynamic modalities while handling noise and unobserved data in order to detect anomalies, track system components, extract relevant trends, modify system controls, or make other informed decisions and accurate inferences about the processes of interest. Dynamic Bayesian networks (DBNs) represent one class of methods that have been successful in such endeavors. A DBN is a graphical model that represents the stochastic processes of a system by encoding domain knowledge of its components and states as a set of variables and their conditional dependencies across time in order to accurately model a time-evolving system while managing uncertainty in the dynamics. Despite their popularity and success, existing toolkits for DBNs generally assume that the networks consist of purely discrete or purely continuous variables, whereas many systems are more appropriately modeled as hybrid DBNs consisting of both types of variables. To address this inadequacy, we develop a Python implementation for hybrid DBNs that incorporates guard functions with Kalman filters and particle filters to support mixtures of discrete and continuous dependencies. We discuss current progress, ongoing work, and future developments and applications of our implementation.

Joint work with Brenda Ng, Lance Bentley Tammero, Goran Konjevod, and Siddharth Manay (LLNL)
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-801817.

37. Using Machine Learning to Compare Simulated and Observed Arctic Climate Data
[S] J. Jake Nichol, Sandia National Laboratories

The extent of sea ice in the Arctic has been declining for decades. Coupled physics-based models (CPMs) forecast less Arctic decadal sea ice extent (km2/decade) than observed in data, leading to conservative predictions. It is important to identify why CPMs are too conservative so that we can revise the CPMs and effectively plan for our future climate. To do this, we train machine learning models on observed data and simulated data separately, then compare the feature importances between the different models. The approach identifies areas in which the CPMs place different emphasis on features or variables from the observational model, which may indicate candidate areas for revision. Observed data comes from satellite sea ice concentration data, along with atmosphere and ocean reanalysis products, in contrast to the simulated data that is generated from the DOE's Energy Exascale Earth System Model (E3SM). For both data sets, machine learning models are fit to predict the minimum sea ice extent in September. Input features from prior months are air temperatures, solar radiation, sea surface temperature, surface pressure, and wind speeds. The major contribution of this work shows that feature importances in the E3SM model are inconsistent between runs and inconsistent with the observed data.

Joint work with Matthew Peterson (Sandia National Laboratories), Kara Peterson(Sandia National Laboratories), David Stracuzzi (Sandia National Laboratories). *SAND2020-0538 A*

38. Designing for Large Simulations: A Case Study
[S] R. Vincent Paris, Iowa State University

We will present an overview of using a large group screening experiment for a deterministic large scale simulation written by the Department of Energy. The simulation had up to 4.5k variables with special interest in second order interactions and good performance on supercomputers. The advantages of the implementation of the group screening design compared to the original analysis will be discussed.

Joint work with George Ostrouchov (Oak Ridge National Lab) Drew Schmidt (Oak Ridge National Lab).

39. Self-Organizing Maps and Their Applications to Data Analysis
[S] Ravi Brannon Ponmalai, UC Irvine

Self-Organizing Maps(SOMs) are a form of unsupervised neural network that are used for visualization and exploratory data analysis of high dimensional datasets. Our goal was to understand how we can use a SOM to gain insights about datasets. We do this by first understanding the initialization, training, error metrics, and convergence properties of the SOM. Next we discuss the ways to interpret and visualize a Self-Organizing Map. Finally we used real datasets to understand what the Self-Organizing Map can tell us about labeled and unlabeled data. Based on experiments with our datasets we found that the Self-Organizing Map can tell us about the spacing and position of high dimensional clusters, help us find non-linear patterns, and give us insight into the shape of our data.

Joint work with Chandrika Kamath, Lawrence Livermore National Laboratory. LLNL-ABS-800086 This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC.

40. Using Anomaly Detection for Activity Identification
[S] Gabriel A. Popoola, Sandia National Laboratories

Distinguishing different types of activities from one another and from the background in sensor data is a fundamental problem in pattern-of-life analysis. We present an unsupervised approach to detecting unique activities in time series sensor data. Our approach looks at changes in the amplitude and frequency distributions of the data over time. We apply methods utilizing Kullback-Leibler divergence and the chi-squared distance metric to detect changes in the time series as well as determine whether two activities are the same type. We evaluate our methods on synthetically generated data.

Joint work with David Stracuzzi, Matthew Peterson. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

41. Uncertainty Quantification for Parallel Discrete Event Simulation
[P] Kevin Quinlan, Lawrence Livermore National Laboratory

Parallel Discrete Event Simulation (PDES) is a useful framework for agent-based models (ABM) and other systems where complex interactions occur. These simulations mimic the operation of a real or proposed system, such as the day-to-day operation of the stock market, the running of an assembly line in a factory, or the interactions on computer networks. PDES may act in non-contiguous and non-linear manners, making traditional response surfaces difficult to use. PDES often include a large amount of stochasticity, which generates heteroscedastic error in the quantities of interest. Often the simulation models a large number of agents with multiple parameter settings each, but the agents may also fall into homogenous classes. Considering these many challenges, we present methodology related to the design and analysis of PDES for multiple applications.

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-766659

42. Scientific Machine Learning using Astrophysical Simulations
[P] Nesar Ramachandra, Argonne National Laboratory

Synthetic data generated via meticulous experimental design, robust numerical techniques and extensive physical validation can be utilized in training machine learning algorithms and statistical applications in cosmological analysis. Such schemes not only assist in meeting the requirement of large amount of representative data in training deep neural networks and compliment in the data space where observed data is insufficient or biased, but also demonstrate a higher degree interpretability and controlled systematics in complex inference algorithms. We narrate the design and utilization of simulated data in three different astrophysical applications, i.e., the galaxy-galaxy strong lensing modeling, cosmological parameter calibration using Gaussian Process emulators, and photometric redshift estimation of galaxies in astronomical sky surveys. In each of these astrophysical applications, the data is generated with physically motivated forward models with carefully chosen priors for the input parameters, and emphasis is given on quantifying prediction uncertainty. We highlight the advantage of using realistic fabricated data in parameter inference problems over real data.

43. Detecting Forest Thinning at the Valles Caldera using Anomalous Change Detection on Synthetic Aperture Radar Data

[P] Elena C. Reinisch, Los Alamos National Laboratory

Forest destruction, whether through human activities or natural disasters, has increasingly become a matter of global concern in recent years. It has been shown that remote sensing is ideal for monitoring deforestation on a broad scale; however, more subtle types of deforestation such as thinning and brush removal are also of interest. We explore the capability of radar remote sensing to detect subtle forest destruction. Radar has the ability to penetrate through tree canopies and capture backscatter information from both the canopy and material beneath it. In particular, we work with synthetic aperture radar (SAR), which can capture such source signatures at fine spatial resolution. We apply anomalous change detection (ACD) to SAR data to detect forest thinning at the Valles Caldera, located in the Jemez Mountains of northern New Mexico. By applying the ACD methodology across dimensions other than temporal, we establish baselines for change detection. Application of ACD across different polarizations also highlights anomalous relationships associated with different types of scattering mechanisms. We also demonstrate the effectiveness of new metrics for distinguishing between anomalies consistently present in data over time and more subtle changes which may be obscured by these anomalies. This is especially useful for analyzing SAR backscatter intensity, which can be dominated by topographic features that are not of interest.

Joint work with Amanda Ziemann (Los Alamos National Laboratory), James Theiler (Los Alamos National Laboratory), Geoffrey Fairchild (Los Alamos National Laboratory).

44. Multi-modal Remote Sensing of Conflict

[P] Christopher X. Ren, Intelligence and Space Research Division, Los Alamos National Laboratory

The detection and quantification of conflict through remote sensing modalities represents a challenging but crucial aspect of human rights monitoring. In this work we demonstrate how utilizing multi-modal data sources can help build a comprehensive picture of conflict and human displacement, using the Rohingya conflict in the state of Rakhine, Myanmar as a case study. We show that time series analysis of fire detections from the Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) can reveal anomalous spatial and temporal distributions of fires related to conflict. This work also shows that Synthetic Aperture Radar (SAR) backscatter and coherence data can be used to detect the razing and burning of buildings and village. Finally we combine the aforementioned modalities with multispectral data to build a multi-modal model to quantify spatio-temporal aspects of the Rohingya conflict.

45. Embracing Unidentifiability in Bayesian Model Calibration with Modularization

[S] Kellin Rumsey, Sandia National Laboratories & UNM

Bayesian model calibration has become a powerful tool for the analysis of experimental data coupled with a physics-based mathematical model. The forward problem of prediction, especially within the range of data, is generally well-posed. However, there are many well-known issues with the approach when solving the inverse problem of parameter estimation, especially when the calibration parameters have physical interpretations. Unidentifiability is at the crux of these issues. In many practical applications, there are a small number of parameters which are considered ``of interest''. By focusing our efforts on these parameters and forfeiting the ability to learn about other parameters, robust inferential procedures can sometimes be obtained via a pseudo-Bayesian approach which is referred to as modularization.

In this article, we present modularization as a general estimation framework and provide a thorough discussion of when the framework should be used and, of equal importance, when it should not. We also develop an efficient algorithm for approximation of the modularization posterior numerically. Using this algorithm, modularization is applied to two example problems - a synthetic dataset based on the borehole function and a dynamic material property application - and the results are compared to the fully Bayesian approach. We show that modularization has many desirable statistical properties when the mathematical model is misspecified, at the cost of conservatism in parameter estimation.

Joint work with Gabriel Huerta (SNL), Justin Brown (SNL), Lauren Hund (SNL). Funding statement: Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

46. Machine Learning for Quantum Computing and Efficient Quantum Computation
[S] A.M. Santana, Oak Ridge National Laboratory

We use a Restricted Boltzmann Machine (RBM) on a Quantum Annealer to perform image classification of particle accelerator data. By encoding the accelerator outputs as the probability distribution of a wave function, the RBM is trained to classify the images by particle type of electron vs positron. We also develop a Deep Reinforcement Learning (DRL) model that optimizes logic gate operations on quantum devices. By decomposing the quantum gates as discretized control pulses, DRL is used to construct fields of finite duration that maximize the precision of the gate operation. This may ultimately be used to develop more time-efficient and noise resilient control pulses for use on quantum computers.

Joint work with A.J. McCaskey.

47. Detecting temporal anomalies in multispectral remote sensing imagery: a case study with Burning Man
[S] Hope W. Simonoko, Los Alamos National Laboratory

The Reed-Xiaoli (RX) detector is used for identifying spatial anomalies in multispectral imagery, which are pixels whose spectra are anomalous relative to other pixels in a scene. The distribution of the spectra in an image is used to represent the background, and the anomalies are the pixels whose spectra deviate statistically from this distribution. While RX is used to identify spatial anomalies, in this research we have instead developed a method to capture temporal anomalies, or fleeting changes, such as a music festival in the desert. Using the annual Burning Man festival as a test case, we use a time series of multispectral images and iterate through each pixel, drawing the ‘background’ distribution from a particular pixel location over time. Temporal RX thus compares a pixel against itself through time, which enables us to capture normal seasonal trends and identify fleeting changes. Using a clustering method, we investigate the nature of the changes detected by Temporal RX to infer types and causes of change.

Joint work with Amanda Ziemann (LANL), James Theiler (LANL), Eric Flynn (LANL), and Geoffrey Fairchild (LANL).

48. Measuring Uncertainty in Object Detection
[S] Mary Angelica Tursi, University of Illinois

Convolutional neural networks (CNN) trained to detect objects are often thwarted by different kinds of noise in the real world. The interaction of this noise with the detection reliability is poorly understood and may produce unexpected effects as seen with adversarial images. For this project, we trained a CNN to detect vehicles in satellite images that were free of noise. We then applied noise in the form of low-pass and high-pass filters and examined how the model's performance in detecting images with confidence is affected when the images are increasingly blurred or sharpened, respectively, generating corresponding 'certainty curve' for the vehicles in the images. Our analysis revealed that detection certainty for each potential object is driven by the energy levels in the mid-range frequencies of the Fourier transform power spectrum for the object. These critical frequencies are individual to each object and correspond to the energy distribution within the power spectrum. Cars whose mid-range energy levels fall at higher frequencies experience a faster drop in certainty level with increasing blur while a more compact energy distribution results in an object detection more resistant to blur. Cars without significant energy outside of the lowest frequencies are poorly detected to begin with. Conversely, cars with a broader spread of initial energy levels are able to withstand a high-pass filter for longer than the more compact energy distributions. Some badly performing cars become temporarily easier to detect when modified by a high-pass filter, which suggests that the CNN is highly sensitive to the mid-range energy level distribution in the object power spectrum.

49. Symmetric Non-Negative matrix factorization with PAC for estimating number of clusters
[S] Raviteja Vangara, Physics and Chemistry of Materials, Theoretical Division, Los Alamos National Laboratory

Of late data-driven modelling has become an important approach in the scientific analysis, and majority of world's data is unlabeled. Unsupervised machine learning is employed to extract hidden patterns in unlabeled data. Clustering is one of such techniques to extract patterns in unlabeled data based on a similarity measure, hence occupies important place in Data Mining. Most of the clustering techniques (like Kmeans) uses distance as the similarity measure. If the data samples are to be clustered not based on spatial distance metric, very few techniques like spectral clustering, which is based on graphical approaches and DBSCAN, which groups based on the density of the vectors, are employed. A prior knowledge on the number of clusters is used in these techniques. Recently, Symmetric Non negative matrix factorization(Symm NMF) is known to be identical to these graph-based approaches, and was demonstrated that it is better than graphical based spherical clustering techniques. To estimate true number of clusters in these non-euclidean data, we employ Proportion of Ambiguously Clustered pairs (PAC), which involve consensus clustering in conjunction with symmetric nmf, to identify number of clusters in the data sets in which the traditional clustering and silhouettes fail.

Joint work with Gopinath Chennupati(Information sciences, CCS-3 ,LANL) and Boian Alexandrov (Physics and Chemistry of Materials, Theoretical Division (T-1) , Los Alamos National Laboratory).

50. Scaling DNA-SIP metagenomics: development of a standardized experiment and analysis framework
[P] Dariia Vyshenska, Joint Genome Institute

Linking specific microorganisms to their functions is one of the fundamental goals of microbial ecology, however, the inability to isolate and cultivate most environmental microbes is a major challenge. The recent development of stable isotope probing (SIP) methods provides an alternative approach for culture-independent identification of microbial populations performing a specific metabolic activity in complex ecosystems. In a SIP experiment, a substrate of interest labeled with a heavy isotope (e.g., 13C, 18O) is incubated with a microbiome sample, where members of the microbial community responsible for the utilization of the substrate will incorporate heavy isotopes in their nucleic acids. DNA (for DNA-SIP) or RNA (for RNA-SIP) can then be extracted and members of the community whose nucleic acids are enriched with heavy isotopes are identified. This approach allows pinpointing active microbial populations even with low abundance. While SIP has traditionally relied on 16S PCR amplicons to identify members of the community enriched in the labeled fraction [1], recent advances in sequencing technologies and bioinformatics tools allow the combination of DNA-SIP and metagenomics, the study of the bulk genomic DNA recovered directly from environmental samples. This enabled the linkage of functions and metabolic reconstructions to (near-)complete microbial genomes from complex communities (instead of single marker genes) [2-5]. Community interest in DNA-SIP metagenomics is quickly rising, however, the lack of standardized protocols and analysis pipelines makes it impractical to apply this technique in high-throughput at present time. To tackle this challenge, we developed a standardized DNA-SIP metagenomic experiment and analysis framework that can scale the adoption of DNA-SIP metagenomics. We demonstrate that the addition of isotopically-labeled synthetic spike-ins for pre- and post-fractionation facilitates quality control, normalization, quantification, and detection of poorly handled samples. With a fully integrated pipeline, researchers will be able to automatically quality control SIP metagenome data, extract a combined set of de novo reconstructed genomes from a single experiment, robustly quantify abundances of these genomes across fractions and samples, and automatically derive enrichment statistics reflecting activity levels of individual microbial populations.

Joint work with Andy Tomatsu, Kanwar Singh, Simon Roux, Rex Malmstrom, Emiley Eloe-Fadrosh. Affiliation: Joint Genome Institute.

1. Saidi-Mehrabad A, He Z, Tamas I, Sharp CE, Brady AL, Rochman FF, Bodrossy L, Abell GC, Penner T, Dong X et al: Methanotrophic bacteria in oilsands tailings ponds of northern Alberta. The ISME journal 2013, 7(5):908-921.

2. Ziels RM, Sousa DZ, Stensel HD, Beck DAC: DNA-SIP based genome-centric metagenomics identifies key long-chain fatty acid-degrading populations in anaerobic digesters with different feeding frequencies. The ISME journal 2018, 12(1):112-123.

3. Eyice O, Namura M, Chen Y, Mead A, Samavedam S, Schafer H: SIP metagenomics identifies uncultivated Methylophilaceae as dimethylsulphide degrading bacteria in soil and lake sediment. The ISME journal 2015, 9(11):2336-2348. 4. Verastegui Y, Cheng J, Engel K, Kolczynski D, Mortimer S, Lavigne J, Montalibet J, Romantsov T, Hall M, McConkey BJ et al: Multisubstrate isotope labeling and metagenomic analysis of active soil bacterial communities. mBio 2014, 5(4):e01157-0111

4. 5. Chemerys A, Pelletier E, Cruaud C, Martin F, Violet F, Jouanneau Y: Characterization of novel polycyclic aromatic hydrocarbon dioxygenases from the bacterial metagenomic DNA of a contaminated soil. Applied and environmental microbiology 2014, 80(21):6591-6600.

51. Statistical Support for Design of Experiments, Quality Control and Process Capability Assessment, and Data Analysis in a Uranium Fuel Fabrication Process

[S] Stephen J. Walsh, Pacific Northwest National Laboratory

The Applied Statistics and Computational Modeling Group at Pacific Northwest National Laboratory (PNNL) supports the development of a fuel fabrication process for United States High Performance Research Reactors (USHPRRs) under the US DOE's Office of Material Management and Minimization (M3). A key component of this office is the conversion from highly enriched uranium (HEU) fuel to low-enriched uranium (LEU) fuel for five USHPRRs and one critical assembly. The motivation for this conversion in United States is the commitment to lead by example and influence international reactor operators to convert to a fuel that is less attractive to diversion from a safeguard's perspective, but effectively maintains the core mission and needs of reactors. USHPRRs require development and deployment of a complex new fuel type that requires significant effort and cooperation across the DOE complex. The process of manufacturing fuel foils for a specific USHPRR involves the following steps (although not all steps have been performed for each campaign):

1. Input Uranium and Molybdenum feed material whose mixture achieves target U-235 and Mo percentages. 2. Cast material into plates. 3. Cut plates into ingots. 4. Homogenize ingots. 5. Can and Hot roll ingot (flatten ingot). 6. Final anneal package. 7. Decan hot rolled foil 8. Possible Cold roll master foil (flatten again). 9. Clad ingot in zirconium, can and hot roll master foil to bond zirconium. 10. Anneal. 11. Decan hot rolled foil 12. Cut into child foils 13. Cold roll foils to final thickness. 14. Final Anneal. 15. Cut into foils of specific dimension for a given reactor type. 16. Clad fuel foils in Aluminum via Hot Isostatic Pressing (HIPing).

In the past years PNNL has managed or will manage several experimental campaigns to test aspects of this process. These include the MiniPlate-1 campaign (MP), OPTimization casting runs (OPT), and the Plate Demo-1 campaign (PD-STD-1). Over the course of these campaigns, multiple aspects of the production process have been investigated and experimented with in order to inform how to implement a capable process that can produce fuel foil that meets specifications for USHPRRs. In this poster we present an expose of statistical tools and data analysis approaches that have been implemented to support optimizing the production process. These tools include:

1. Variance components estimation and uncertainty analysis. 2. Automated image analysis to measure fuel meat and cladding thickness. 3. Design of Experiments and Response Surface Methodology for assessing Al clad bond strength under different HIP conditions. 4. Design of Experiments to propose blocked split-plot designs for testing levels of four process variables while controlling nuisance factors.

Joint work with Deb K. Fagan, Kevin Anderson, Scott Cooley, Alejandro Heredia-Langner, Dan Hodges (PNNL).

52. Comparing Unfiltered Topic Analysis Approaches to Understanding Social Unrest in Ukraine

[S] Chrysm Watson Ross, Los Alamos National Laboratory/UNM

We build a set of natural language processing (NLP) topic models for predicting political protests without any pre-filtering for 'significant' terms. We compare several approaches to predict real time political protest behavior, and determine that newer 'deep' embedding models can perform better than classic vector space models (VSMs.) This proves the utility of embedding models in inference applications on high speed open source data, as they don't require the extensive pre-processing overhead of VSMs. We additionally extend emoji parsing and rare language libraries to the interpretation of short texts in an uncommon language (tweets in Ukrainian.) We apply these approaches to the 2013-2014 Euromaidan protests in Kyiv, Ukraine as a test case.

53. Multi-Model Penalized Regression for Feature Selection

[S] Laura Wendelberger, North Carolina State University

Penalized regression is a useful tool in feature selection, but many methods fail to take into account the existence of model uncertainty. In a dataset with highly correlated predictors, as is the case in predicting materials properties from structural information, a single interpretation of the model identifies only one set of possible covariates when in reality there may be several. We propose a Multi-Model Penalized Regression (MMPR) to identify several possible linear models which differ from each other with respect to the presence of covariates and their corresponding coefficients. We introduce a possible form for a model similarity penalty which may encourage variable shrinkage and/or sparsity among the different models based upon the settings.

Joint work with Brian Reich (North Carolina State University), Alyson Wilson (North Carolina State University).

54. Robust Detection of Auxiliary Equipment for Autonomous Facility Resource Management
[S] Brian M. West, Los Alamos National Laboratory

Large scale operations such as machine shops, laboratories, or industrial plants can have dozens or hundreds of different machines operating throughout a day. Management of the operation of these machines can be difficult, particularly when manned by few individuals and processing steps in a facility are interconnected and depend on one another. If one or more machines have stopped due to a malfunction, a lack of resources for machines further in the process can lead to inefficiencies or could lead to further malfunctions. A method demonstrated in this study is to evaluate the state of the machines simultaneously by monitoring the electromagnetic signatures emitted from the auxiliary components of the machines on the power line feeding the facility. Combining known signatures from testbed databases and physics-based knowledge of machine operation, robust detectors and classifiers can be developed to identify what components are functioning and determine machines are operating or idle. This information can then provide a real-time operational health status of a facility with a limited number of required sensors.

Joint work with Eric B. Flynn, Los Alamos National Laboratory.

55. Applying Autoencoders for Model Reduction of Multigroup Scattering Matrices
[S] Ben Whewell, University of Notre Dame

The amount of memory needed to store the nuclear data required to perform a neutron transport simulation is becoming a limiting factor with advanced computing platforms. In this poster, we apply autoencoders, a form of unsupervised learning using neural networks, to reduce this memory requirement. We find that autoencoders are able to represent the results of the scattering matrices for different temperatures accurately through the use of an L2 and L-infinity loss function. This reduction of data minimally affects the results from transport calculations when compared with results using the full, original data. The results from the autoencoder can also give us insight into the true dimensionality of the scattering data.

56. A new Deep Learning architecture for general disruption prediction across tokamaks
[S] Jinxiang Zhu, MIT PSFC

Nuclear fusion power delivered by magnetic-confinement tokamak reactors hold the promise of sustainable and clean energy. The avoidance of large-scale plasma instabilities called disruptions within these reactors is one of the most pressing challenges, because disruptions can halt power production and damage key components. Predicting plasma instabilities and disruptions from first-principle models has proven to be extremely difficult because of the inherent complexity of the problem. On the other hand, recent statistical and classical machine-learning approaches based on real-time measured data have shown promising results for disruption prediction. Different tokamak reactors have very different operational space, spatio-temporal scales and diagnostics. Therefore, most of the previous approaches were developed specifically for one reactor and didn't show promising cross-device prediction ability, which will be key for powerful near-future burning plasma machines such as ITER, as they cannot withstand more than a few disruptions. In this poster, we selected a set of physical signals which are available on most of the tokamak reactors, and developed a general disruption prediction algorithm based on Deep Learning using three large databases of both disruptive and non-disruptive discharges from three tokamaks: Alcator C-Mod, DIII-D and EAST. The previous disruption warning databases for these three machines had a different sampling scheme for the disruptive and non-disruptive shots, and they also had different sampling rates. To develop a general prediction algorithm, we interpolated the three disruption warning databases onto uniform sampling rates (different for each machine) based on the flattop length of each machine. The predictor algorithm was trained on flattop data consisting of the same 12 plasma signals on all three machines, with a uniform 5 ms sampling rate on C-Mod, 10 ms sampling rate on DIII-D and 25 ms sampling rate on EAST. A shot-by-shot general testing scheme has been developed to give a disruption warning alarm for all three machines and we successfully achieved state-of-the-art performance on all three test datasets. Moreover, several important guidelines about how to extrapolate the present predictor to future devices will be presented via cross-machine study. And the possible physics explanations of these conclusions will be given.

1. Increasing Fairness and Understanding of Data Competition Leaderboard Scoring
Christine M. Anderson-Cook, Los Alamos National Laboratory

Data competitions, such as the Topcoder contest focused on urban radiation search (https://www.topcoder.com/lp/detect-radiation) recently hosted by Los Alamos National Laboratory, are becoming increasingly common as cost-effective ways of obtaining diverse solutions to difficult problems from different experts. However, a known limitation of these competitions is their dependence on the specific data presented to the competitors and used to compare their solutions. While the dependence of results on the choices made by the competitions hosts can be an important advantage to drive improvements in areas of importance, there is a subjective element to the final rankings that may feel artificially rigid from the competitors' perspective. Often rankings are decided by small differences in the scoring metric that can get shuffled if a slightly different set of data was used, while the prize money amounts for each final ranking can vary substantially. In this poster we propose methods for more fairly distributing the competition prize money that reflects the closeness of competitor solutions as well as numerical and graphical summaries to help the competition host understand the relative performance of the top solutions for the general problem of interest, instead of the particular dataset provided in the competition.

Joint work with Lu Lu, University of South Florida.

2. Physics Based Machine Learning for HPC Computational Design
Rick Archibald, Oak Ridge National Laboratory

Simulation based computational design can quickly become an intractable problem depending upon both the size of the computational design space and computational complexity of design simulation. We propose the use of physics-based machine learning and high performance computing (HPC) as solutions that make computational design possible for complex simulations and design geometries. Our design problem is based upon the optimization of neutronics and thermo-hydraulics simulations, where computational cost only allows for a relatively sparse number of test geometries. We develop reduced order surrogate models for both neutronics and thermo-hydraulics, that can quickly sample hundreds of thousands of geometries on SUMMIT, the GPU based HPC system at ORNL. Using the combination of surrogate modeling and sparse validation and correction from full-physics simulation, we are able to train Gaussian Process (GP) machine learning methods to accurately predict optimal designs.

3. Multiple Object Tracking for Space Traffic Management
Jason Bernstein, Lawrence Livermore National Laboratory

Space traffic management requires the ability to track multiple objects in orbit with high-precision and in clutter. Several techniques have been developed for multiple object tracking, such as the joint probabilistic data association filter (JPDAF), but these methods often make linear, Gaussian motion and measurement error assumptions or approximations, which are not always realistic for space object tracking. For example, the Keplerian orbit model specifies elliptical motion, and measurement data is typically of the range or bearings of the objects with respect to an observer. This poster introduces the multiple object tracking problem, summarizes some of the main statistical techniques for its solution, and considers the applicability of these techniques for space object tracking. Simulation results are provided to demonstrate how a representative method performs under different tracking scenarios.

LLNL-ABS-800393 This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

4. Scale-bridging Gaussian model for high-resolution LIDAR wind data
Julie Bessac, Argonne National Laboratory

Wind conditions are typically measured with various instruments, wind speed is generally collected as averages over a given time-window. However, time-averaged wind speed might not be representative of potential sub-sample variability as for instance wind extremes. In this work, we will discuss a scale-bridging regression for vertical wind speed extracted from DOE-ARM Doppler LIDARs (light intensity and ranging radar) recorded at second/meter levels. We characterize statistically information across different fine resolutions about the space-time structures of wind profiles that are typically misrepresented at larger resolutions. We propose a fine-scale statistical model, based on a linear regression model, that aims to bridge the gap between the available fine resolution data to more commonly used coarser resolutions.

Joint work with Emil Constantinescu (ANL), Mihai Anitescu (ANL).

5. Multi-Stage Emulation and Uncertainty Quantification for Disaggregation of LIBS Spectra

K. Sham Bhat, Los Alamos National Laboratory

Identifying the chemical composition of a target using direct laser induced breakdown spectrography (LIBS) measurements is often difficult. While domain experts regularly can identify the presence or absence of a compound in a spectrum, the relative abundances pose a challenge due to interactions (called 'matrix effects') between atoms in a plasma. Matrix effects are challenging as it changes peak heights in a nonlinear manner; and hence the spectrum for a target is no longer a simple linear combination of the spectra of the individual compounds. Using ATOMIC, a LANL based physics simulation code which predicts spectra under the influence of matrix effects for a specific chemical composition (and other parameters). However ATOMIC is very complex for large compounds with many parameters, which makes traditional emulation difficult. A multi-stage emulator is developed to estimate ATOMIC for a complex compound using much faster single element emulators, which can then be used with Bayesian model calibration techniques to solve the disaggregation problem.

Another challenge is the scale differences between simulation and observation; in this case between the ATOMIC model output spectra and the measured spectra from the ChemCam instrument. A potential solution shown here uses a highly structured discrepancy within a Bayesian model calibration context to capture the systematic scale differences between simulations and observations.

6. Natural Language Processing for Topic Identification supporting Document Search and Identification for Nuclear Proliferation Detection

Jonathan Bisila, Sandia National Laboratories

Open source analysts for the International Atomic Energy Agency's Department of Safeguards use complex search strings to identify candidate open source scientific publications to review as part of their analysis of the cohesion of a state's declared and actual nuclear activities. While these search strings have been refined over time by expert fuel cycle and open source analysts, they are inherently limited by the requirement of term presence. We hypothesize that topic models relevant to the nuclear fuel cycle may be able to grasp a broader set of potentially relevant documents by allowing the inclusion of related (as determined by natural language processing topic models) documents that do not have required terms to be included for analysis.

To evaluate the potential of these methods, our team has developed a workflow to determine how various topic modeling and clustering techniques can be used to identify topics and documents of interest, comparing our results to a traditional analyst search string. In this poster we will define our workflow, describe an initial implementation of the workflow, and present preliminary results.

Joint work with Daniel M. Dunlavy (SNL), Zoe Gastelum (SNL). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2020-0279A

7. A Unique Similarity Metric for Anomaly Detection in Temporal Networks

Katherine Cauthen, Sandia National Laboratories

The ability to detect changes between similar networks can be viewed as an anomaly detection problem. This paper presents a network similarity metric that is sensitive enough to detect subtle changes in link strength or network structure without early saturation when dramatic changes must be detected. The algorithm achieves this level of fidelity by combining multiple network analysis algorithms in an efficient manner. The anomaly detection methodology when applied to the metric makes use of both distribution fitting and control chart techniques. Validation of the algorithm is accomplished using advanced measurements based on Principal Component Analysis (PCA).

Joint work with Nathanael Brown - Sandia National Laboratories Justin Durfee - Sandia National Laboratories Chris Frazier - Sandia National Laboratories Linda Nozick - Cornell University. Funding Statement: SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525. SAND2020-0627 A.

8. Systems abstraction in XMILE for automated dynamic system interpretation

Jackson L. Chin, Pacific Northwest National Laboratory

Dynamic system models are powerful tools for approximating complex systems, and their interpretations can offer insights into opaque processes. A key limitation in these interpretations is in the diversity of model subjects and formats. Dynamic systems span a multitude of disciplines and theoretical bases, and it is difficult to implement one underlying framework that can reliably reproduce all dynamic systems. Without a standardized means of representing such systems, it is difficult to develop tools that can consistently interpret different dynamic systems. Here, in an effort to identify a singular format that can represent any dynamic system, we have studied and developed a simulator for XMILE, an interchange language that encodes a dynamic system model within an .xml file. We then apply XMILE to represent and simulate a variety of systems including a Lotka-Volterra predator-prey system and a nuclear reactor. Through these applications, we demonstrate XMILE's ability to abstractly represent a variety of dynamic systems and show how our simulator can accurately recreate models and generate data for system interpretation tasks.

Joint work with Lee Burke, Pacific Northwest National Laboratory Romarie Morales, Pacific Northwest National Laboratory Dave Engel, Pacific Northwest National Laboratory Paul Whitney, Pacific Northwest National Laboratory.

9. Machine learning uncertainty quantification for reduced order models of hypersonic flows

David S. Ching, Sandia National Laboratories

Machine learning methods are used for uncertainty quantification of reduced order model simulations, following on the work of Freno and Carlberg (2019). Reduced order models present a computationally inexpensive alternative to full simulations by projecting the governing equations on a reduced proper orthogonal decomposition basis computed from full order simulations, but the projection can introduce errors in the reduced order model. Reduced order models are developed separately for steady-state hypersonic flows around a sphere and around the HIFiRE-1 vehicle. Errors in the reduced order models are predicted using the simulation parameters and a subset of the governing equation residuals as inputs. The machine learning methods used include neural networks, support vector regression, and random forests, and are compared to more traditional least squares regression.

Joint work with Patrick Blonigan. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2020-0737A

10. Machine Learning Approaches to Identify the Drivers of Performance and to Predict Productivity of Bakken Oil Wells

Tim Coburn, The University of Tulsa, School of Energy Economics, Policy & Commerce

Geologists and petroleum engineers have struggled to identify the mechanisms that drive productivity in horizontal hydraulically-fractured oil wells producing from the middle member of the Bakken Formation. Although, in general, productivity of such wells is thought to be linked to reservoir oil saturation and to conditions of over-pressure, their performance is highly variable, which may add additional uncertainty to U.S. energy projections. To identify the best predictors of the cumulative oil produced in the initial 360 days of oil production, that is, the measure of wells productivity used here, Random Forest and gradient boosting algorithms were applied to a dataset containing more than 7300 horizontal hydraulically-fractured wells drilled into the middle member of the Bakken Formation from 2010 through 2017. The goal is to use data-driven machine learning approaches to identify the best predictors thought to vary spatially with geology and well completion predictors that include the number of fracture stages, volume of proppant per stage, and the volume of injected fluids per stage. The algorithms identified initial gas production, the 180-day water cut, and vertical depth as important predictors that vary spatially. The relative strength of these predictors was compared with well-completion predictors. Well productivity and the predictor data were used to calibrate play-wide and subarea models and demonstrated the variability of the relative importance of the productivity predictors between sub-areas. The analysis further showed the incremental improvements in well productivity resulting from changes in well completion/design variables. In summary, the predictive performance of the machine learning approaches suggests that they may be effectively applied to predict the productivity of wells at undrilled sites in the Bakken play, which could ultimately lead to much improved estimates of future U.S. energy resources.

Joint work with Emil Attanasi, US Geological Survey (1st author) Phil Freeman, US Geological Survey (2nd author)
I am the 3rd author.

11. Monotone CDF estimation for binary data via kernel regression and evolution
David H. Collins, Los Alamos National Laboratory

Binary response data, encountered in applications ranging from market research to toxicology, are typically analyzed using logistic or probit regression, which imply latent threshold variables subject to parametric models (logistic or Gaussian, respectively). These can produce misleading results if the true threshold model does not match assumptions. The 'pool adjacent violators' method [1] provides a nonparametric maximum likelihood estimate (MLE) for the threshold distribution (CDF), which minimizes the sum of squared errors relative to the observed data. We offer an alternative, based on kernel regression followed by an evolutionary algorithm that guarantees a monotone estimate of the CDF. Based on simulations, we conjecture that this method minimizes the sum of squared errors relative to the true CDF.

[1] M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955). 'An empirical distribution function for sampling with incomplete information,' The Annals of Mathematical Statistics, 641-647.

12. nPINNs: nonlocal Physics-Informed Neural Networks
Marta D'Elia, Sandia National Laboratories

Nonlocal models provide an improved predictive capability thanks to their ability to capture effects that classical partial differential equations fail to capture. Among these effects we have multiscale behavior (e.g. in fracture mechanics) and anomalous behavior such as super- and sub-diffusion. These models have become incredibly popular for a broad range of applications, including mechanics, subsurface flow, turbulence, plasma dynamics, heat conduction and image processing. However, their improved accuracy comes at a price of many modeling and numerical challenges. In this work we focus on the estimation of model parameters, often unknown, or subject to noise. In particular, we address the problem of model identification in presence of sparse and possibly noisy measurements. Our approach to this inverse problem is based on the combination of 1. Machine Learning and Physical Principles and 2. a Unified Nonlocal Vector Calculus and Versatile Surrogates such as neural networks (NN). The outcome is a flexible tool that allows us to learn existing and new nonlocal operators. We refer to our technique as nPINNs (nonlocal Physics-Informed Neural Networks); here, we model the nonlocal solution with a NN and we solve an optimization problem where we minimize the residual of the nonlocal equation and the misfit with measured data. The result of the optimization are the weights and biases of the NN and the set of unknown model parameters. We provide several numerical results that illustrate our findings, including an application to turbulence.

13. Hindcasting Violence in Colombia Using Data Fusion Methods
Ashlynn R. Daughton, Los Alamos National Laboratory

Colombia experienced a decades-long civil war between the government and many left-wing guerrilla groups. It was marked by violence, kidnappings, and large quantities of human displacement. Monitoring and forecasting civil wars are important to mitigate their potential impact but require access to ground truth data. We examine the use of Internet data streams, namely Google search queries, tweets related to politics, and traditional news sources to retrospectively forecast (i.e., hindcast) state-based armed violence in Colombia. We compare the results of statistical models using three combinations of these features to evaluate the predictive capabilities of each data stream. Our results show that the combination Internet and traditional news data models perform most consistently, though Internet-only is surprisingly promising. Overall, we are able to produce high-quality models hindcasting the presence or absence of state-based armed violence in Colombia up to 6 months in advance. These results support the use of exogenous data streams to forecast evolving situations around the globe.

Joint work with Sara Del Valle, Chrysm Watson Ross, Geoffrey Fairchild.

14. Towards in-situ inference for climate extremes
Mary Frances Dorn, Los Alamos National Laboratory

We present initial groundwork towards performing statistical inference in-situ (implemented within HPC simulation codes) for climate extremes. This work explores the spatial variability of temperature extremes using Bayesian inference to fit extreme value distributions on data from E3SM, DOE's Earth science model. Additionally, we examine the effect of priors for extremes on parameter estimates as well as derived quantities of interest such as return periods. This is all done in the Julia scientific computing and data science language, which is built to support the high performance needs of in-situ inference.

15. Detecting Flaws in Software Programs using Multimodal Deep Learning
Daniel M. Dunlavy, Sandia National Laboratories

Efficient, reliable, hardened software plays a critical role in a wide spectrum of DOE mission areas. Current practice of characterizing and analyzing software for these important mission applications is often limited to a narrow view of a single functional representation of a program; e.g., source code, binary, or disassembled code fragments. Moreover, much of this characterization is performed in manual assessments, resulting in slow and laborious processes. Different program representations contain different information, and what exactly is lost between representations is often not identified or quantified, leaving software analysts to reason about program correctness and potential flaws based on incomplete information. We explore the use of multiple deep learning models for detecting flaws in software programs. Current, standard approaches for flaw detection rely on a single representation of a software program (e.g., source code or a program binary). We illustrate that, by using techniques from multimodal and crossmodal learning, we can simultaneously leverage multiple representations to improve flaw detection over single representation analyses. Much of the work in multimodal and crossmodal learning has focused on recognition and prediction problems associated with two or more representations of transcript, audio, video, or image data. We aim to leverage those methods and results to develop an improved software flaw detector that learns from both source code and binary representations of software programs. Specifically, we adapt three deep learning models from the multimodal learning literature for use in flaw detection and demonstrate how these models outperform traditional deep learning models. We present results on detecting injected flaws in the Linux Kernel. Our contributions include the following: - The first application of multimodal learning for software flaw prediction (as far as we are aware); - A comparative study of three deep learning architectures for multimodal and crossmodal learning applied to software flaw prediction; and - A data set of software flaws with alignment across source code and binary function instances that can be used by the multimodal and crossmodal learning research communities for benchmarking new methods.

Joint work with Scott Heidbrink (SNL), Kathryn N. Rodhouse (SNL). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2020-0755A

16. Characterizing transfers of nuclear material through multi-sensor indicators
Adin Egid, Los Alamos National Laboratory

Identifying when a transfer of nuclear material has occurred and characterizing the type of transfer is an important non-proliferation task. Individual sensors can indicate informative features about the transfer, but a more complete picture can be obtained when intelligently combining indicators from multiple modalities. This work will present the fusion of indicators obtained from radiation and acoustic signals placed around a nuclear facility, including a research reactor and reprocessing facility at Oak Ridge National Laboratory. Analysis of total gamma counts from an outside radiation network determine the movement of the nuclear material while acoustic signals indicate a vehicle used to perform the transfer. Combining these pieces of information will allow us to determine evidence for a particular type of transfer being performed. We will also demonstrate how informative features from acoustic data can be identified through a joint analysis with the radiation data and present an anomaly detection approach to determine when a transfer has occurred.

Joint work with Dave Osthus (LANL), Brian Weaver (LANL), Jon Woodring (LANL), Emily Casleton (LANL), Dan Archer (ORNL), Andrew Nicholson (ORNL), Milton Garces (UHawaii), Edna Cardenas (INL), Will Ray (ORNL), and Tom Reichardt (Sandia).

17. Asynchronous versus Bulk Synchronous Many-to-Many Long Read Alignment
Marquita Ellis, UC Berkeley and Lawrence Berkeley National Laboratory

Third generation DNA sequencing offers enormous advantages for biological analysis and insight, however current sequencing instruments emit errors at high rates and therefore analysis of the output long reads requires different approaches than their short read counterparts. In previous work, we presented the first distributed memory overlapper and aligner specifically designed for long reads and parallel scalability, implemented in the bulk synchronous pipeline, diBELLA [1]. We improve upon this work by investigating approaches for achieving space lower bounds while simultaneously minimizing the communication overhead of the distributed many-to-many alignment. We highlight the advantages of asynchronous algorithms, implement a baseline asynchronous many-to-many alignment algorithm in UPC++, and demonstrate its scalability on NERSC's Cray XC40, Cori, with both the Intel Knight's Landing and the Intel Haswell node architectures.

[1] M. Ellis, G. Guidi, A. Bulu', L. Oliker, and K. Yelick. diBELLA: Distributed long read to long read alignment. In ICPP 2019, Proceedings of the 48th International Conference on Parallel Processing, (Kyoto, JP)

18. Compound Refractive Lens Alignment using Laser Intensity-Adapted Stochastic Gradient Descent
Malena Español, Arizona State University

Compound refractive lenses (CRLs) are used to focus X-rays and are typically manually aligned in-situ at facilities such as synchrotrons and X-ray free electron lasers (XFEL). This time-intensive process interferes with limited beam time. As XFEL facilities tend to have fairly unstable beams, the intensity of the beam fluctuates, affecting calibration of the CRLs. Numerical techniques can be used to drive calibration, but standard optimization methods, such as gradient descent, are insufficient due to the beam jitters. We correct for the wandering laser intensity by adapting stochastic gradient descent to include a novel sampling method and demonstrate its effectiveness on simulations of experimental conditions at the Advanced Photon Source beam line.

Joint work with Sean Breckling (Nevada National Security Site), Eric Machorro (Pacific Northwest National Laboratory), Michael Brennan (MIT), Eric Clarkson (University of Arizona), Ryan Coffee (LCLS-SLAC National Accelerator Lab), Jessica Pillow (University of Arizona), Ajana' Williams (Nevada National Security Site), Marylesa Howard (Nevada National Security Site).

19. Multi-Modal Event-Based Monitoring of Radiochemical Processing
Garrison S. Flynn, Los Alamos National Laboratory

The Radiochemical Engineering Development Center (REDC) at Oak Ridge National Laboratory is a multipurpose radiochemical processing and research facility that is home to a variety of laboratory space. One of the key missions supported by REDC is production of radionuclides, including Pu-238 and Cf-252. Target rods are manufactured at REDC and irradiated at the High Flux Isotope Reactor for isotope production. Once irradiated target rods return to heavily shielded hot cells at REDC, they undergo a series of dissolution processes to recover the isotopes of interest. In this project, we present an integrating framework to provide context for measurements of independent sensing modalities to detect and identify dissolution events of interest. We combine effluents measured using a high purity geranium (HPGe) detector in the REDC ventilation stack with electromagnetic signals collected using current transformer (CT) clamps outside the facility. Through inverse analysis a timeline of effluent and electromagnetic observations are used to determine likely parameters of the dissolution process.

Joint work with Christopher X. Ren, Nidhi K. Parik: LANL Dan Archer, Riley Hunley, Tom Karnowski, Andrew Nicholson, Randall Wetherington: ORNL.

20. NukeBERT: A nuclear domain BERT model for improved natural language processing tasks in nuclear nonproliferation
Daniel Fortin, Pacific Northwest National Laboratory

Natural language processing (NLP) tasks (text classification, entity recognition, etc.) have seen amazing improvements over the last few years. This is due to word embedding models such as BERT that achieve deep knowledge transfer by using a large pretrained model, then fine-tuning the model on specific tasks. Some have called 2019 the year of BERT. The BERT architecture has shown even better performance on domain specific tasks when the model is pretrained using domain relevant texts. Inspired by these recent advancements we are developing NukeBERT, a nuclear-domain BERT model pretrained on 1.5 million abstracts from the DOE Office of Scientific and Technical Information (OSTI) database. The NukeBERT model will be utilized by downstream tasks that are part of investigations under this project (e.g. classification of research articles, supplying edge-weights in graph analytics, filtering large trade databases based on focused topics). We describe the NukeBERT model and highlight its potential use in these various downstream tasks. In general, the NukeBERT model can be fine-tuned for specific nuclear nonproliferation NLP tasks and can be utilized by other DOE projects by making the pretrained model available on an eOS Azure instance.

Joint work with Ben Wilson, Rustam Goychayev, Karl Pazdernik, Lisa Newburn.

21. A minimally supervised data fusion and classification method for event detection
Chris Frazier, Sandia National Laboratories

We demonstrate a method for detecting relatively rare events of interest from (potentially correlated) multivariate data such as output from multiple detectors. The method consists of (1) a modified ensemble method for novelty detection which nominates potential events for Subject Matter Expert (SME) review, (2) SME tagging of the nominated cases combined with automated negative tagging of the least-novel cases from the ensemble method, and (3) data fusion and classification with Bayesian Networks. The modified ensemble method allows fine control over the amount of data the SME must review. Additionally, using a Bayesian Network for classification allows detecting multivariate patterns in the positively-tagged cases, enables partial learning (i.e., of the raw multivariate relationships) even from non-tagged cases, and natively provides likelihood estimates for classifications. Both the ensemble method and the Bayesian Network are explainable 'clear box' methods whose results can readily be examined to understand why certain scores or classifications were provided.

SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.

Joint work with Nathanael Brown, Sandia National Labs Matt Hoffman, Sandia National Labs.

22. Extracting Stable Community Information on Relational Data
Kasimir Gabert, Sandia National Laboratories

Many datasets critical to national security require not only individual data points but also, and more often more importantly, relationships among them, resulting in a natural representation as graphs. In many cases, these graphs are built out of constantly changing real-world data and become unwieldly large, resulting in both computational and analytical challenges. A giant hairball may make a nice presentation background image, but it won't lead to useful insight. Despite their global sparsity (e.g., the likelihood of two random people knowing each other is low), many graphs representing real-world phenomena, contain locally dense regions. These dense regions play a significant role in extracting information from the graph, enabling computation on it, and developing visualizations for it. Our work extends dense subgraph computation from static graphs to dynamically changing graph streams and then applies this to the problem of uncovering and discovering communities.

The densest possible subgraphs are fully connected graphs, known as cliques, which are computationally intractable to find. Recent work has moved to finding cores of graphs instead, which are found in linear time and are connected subgraphs with a bound minimum degree. We focus on finding nuclei, which are denser and more useful than cores but still tractable compared to cliques. We develop an algorithm and implementation that operates on changing graph streams, maintaining full nuclei decomposition hierarchies. This algorithm is faster than re-computing from scratch and is able to handle concurrent insertions and deletions at high data rates.

We then apply this algorithm in two ways: first, we directly run it on real-world, changing graphs to keep track of dense hierarchies. Such hierarchies form trees where the densest regions are the parents of less dense, but larger, regions. We are studying whether such hierarchies form interesting communities, and how tracking them over time can yield insight into community evolution. Second, we have initial work-in-progress on using these hierarchies as a guide to label the output of other well-established community detection algorithms. Such a labeling effectively turns static community detection algorithms into dynamic algorithms and could be used to understand the stability and behavior of the algorithm over time.

We developed a strategy to maintain graph nuclei on dynamic graph streams, enabling nuclei's use on real-world changing graphs. Our implementation is developed for general nuclei and faster than other highly specialized maintenance algorithms, including for cores. Maintaining nuclei both directly results in an interesting and useful hierarchy and also may prove to be a crucial tool for evaluating and understanding existing community detection algorithms.

Sandia National Labs is managed and operated by NTESS. a wholly owned subsidiary of Honeywell International for the US DOE's NNSA under contract DE-NA-0003525.

Joint work with Ali Pinar, Sandia National Laboratories.

23. Leveraging a Machine Learning Approach to Solve an Inverse Problem Performance Query
Christopher J. Garasi, Sandia National Laboratories

Statistical, machine learning (ML), and deep learning algorithms can be used to take combinations of multidimensional input parameter values and output performance metrics (e.g. regression data) as training data and create a tuned model. Quantitative assessments of model accuracy can then be made with respect to a separate validation dataset. Once tuned, the model provides users with the ability to estimate output performance for a given set of input values with nearly instantaneous turn-around time. The model solution using causal input parameters to relate to output observations is referred to as a 'forward problem' solution.

In certain circumstances, a user will want to select a desired output performance and determine the valid set of input parameter values which will achieve the specified performance. Model solutions to this 'inverse problem' can be very challenging, depending on the dimensionality of the input and output data. We will present a practical approach to solving the inverse problem workflow by using a model which was created via a ML algorithm for the forward problem. Through a combination of searching algorithms and visualization techniques we will use results from the forward model to obtain and reduce the number of valid input parameter value combinations which will meet the desired output performance query.

Joint work with David Dooling, PhD, RS21, Albuquerque, NM Kameron Baumgardner, MS, RS21, Albuquerque, NM. *Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.*

24. The Use of Synthetic 3D Images to Drastically Reduce Real-World Training Data for Object Detection Models
Zoe N. Gastelum, Sandia National Laboratories

Photos from hand-held cameras that are shared via social networks, blogs, or news media are an increasingly abundant information source, and are part of an all-source approach to nuclear proliferation detection. The quantity of available information makes it unrealistic for analysts to search for potentially-relevant images without assistance. Deep learning object detection models are a promising option for this task but require significant domain-relevant training prior to be usable for this domain. Deep learning object detection models train on thousands to millions of example images to accurately locate and identify the object of interest. These models can have high levels of accuracy - sometimes better than human analysts - at recognizing objects of interest if they are well represented in the training data. Most deep learning models for object detection that are openly available today do not include content relevant for nuclear proliferation detection. Even using a pre-trained model and fine-tuning for proliferation-relevant images requires thousands of images. In the domain of nuclear proliferation detection, that quantity of images might not exist, may be too sensitive to use in model training, or may lack ground truth labels (which are time-consuming and potentially error-prone to assign). In this project, we are developing parameterized three-dimensional computer models of objects of interest to generate unlimited numbers of synthetic two-dimensional images for training. This approach allows us to: -Vary the object of interest within known physical limits, modifying size, shape, material, etc. -Vary the background in which the object is situated, or eliminate the background entirely. -Determine what, if any, false cues the model is learning and generate additional images to counter-balance our training set. Our goal is to experimentally develop an approach which will allow us use synthetic images to train deep learning object detection models, while limiting or eventually excluding real-world images from training, to achieve comparable performance to those models trained on exclusively real-world images.

Joint work with Timothy Shead, Sandia National Laboratories Michael Higgins, Sandia National Laboratories.

25. Data Assimilation for the Radiation Belt Environment using the Four-Dimensional Variational Method
Humberto C Godinez, Los Alamos National Laboratory

In this work we implement a four-dimensional variational data assimilation method (4D-Var) to a radiation belt model in order to determine key model parameters. A particular hurdle in implementing the 4D-Var is the computation of the adjoint model. In our work we derive the continuous adjoint model associated with the radiation belt model and discretize them both with a Crank-Nicholson scheme. The resulting adjoint solution is sufficiently accurate for convergence in the 4D-Var assimilation method. We use the 4D-Var assimilation to determine diffusion coefficients (D_{LL}) within the model that control how particles are distributed throughout Earth's magnetic environment. A series of assimilation experiments are perform to validate the methodology using synthetic data. A more realistic assimilation experiment is performed using Van Allen Probes data for phase space density. The assimilated radiation belt estimate closely match the data by estimating appropriate radial diffusion coefficients for the time-period of the data.

Joint work with Michael G Henderson -- Los Alamos National Laboratory.

26. Identifying and Clustering Activity in Seismic Waveforms

Renee Gooding, Sandia National Laboratories

Abstract: Identifying and Clustering Activity in Seismic Waveforms In this work we describe a data-driven approach for identifying and clustering activities, such as vehicle movement, in seismic waveforms. We leverage established methods for pre-processing raw seismic signals, such as power-spectral-density analysis, and for extracting discrete seismic events, such as enveloping. Using the extracted events, we explore two methods for clustering the events into groups that represent similar activities. The first approach uses wavelet transforms to identify key features of the waveform and then applies k-means clustering. The second method is novel in the seismic domain and bypasses the need for feature engineering by using the normalized compression distance between events to quantify event similarity. Preliminary results for both clustering methods on real sensor data are presented along with a discussion of the technical challenges and implications for each approach. We conclude with directions for further development.

Joint work with Erick Draayer, Nicole McMahon, Dylan Anderson, David Stracuzzi, Sandia National Laboratories. *Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.*

27. Augmenting Expert Search for Sources of Bias in Nuclear Data Validation Benchmarks Using Machine Learning

Mike Grosskopf, Los Alamos National Laboratory

Estimates of cross sections for isotopes in nuclear reaction processes are critical for understanding and modeling nuclear physics in reactors and other scientific applications. These estimates are tested in simulation of validation benchmarks, which integrate many sets of nuclear data into one model of a complex experiment. Differences between the simulation results and experiment feedback into a search for ways to improve the nuclear data estimates. Because the experiments depend on a wide set of nuclear data in complicated, inter-dependent ways, the process is largely driven by expert judgement. Utilizing machine learning tools for prediction and interpretability to learn and communicate complex, high-dimensional relationships can augment this expert search. We present results using random forests to predict the bias in validation benchmarks and using model interpretability metrics for assessing importance in the high-dimensional, highly-correlated feature space.

Joint work with Denise Neudecker, Michel Herman, Wim Haeck, Scott Vander Wiel, Mike Rising, Alex Clark, Pavel Grechanuk (Oregon State University).

28. Extreme values of physical processes and dynamical systems with random initialization

Charlotte Haley, Argonne National Laboratory

Extremes of wind speed can be unpredictable in the presence of complex orography. In this study we examine a model for vertical wind speed that accounts for the non-stationary diurnal fluctuations of wind speed due to convection and gravity wave impacts at the top of the boundary layer and consider hierarchical Bayesian approaches to modeling arbitrary level-crossings based on renewal theory. We additionally address a similar problem of computing extreme excursion probabilities in dynamical system arising from random initialization. These are of interest in power systems, design of critical infrastructure, etc. Accurately computing extreme excursion probabilities help us to balance costs and risks. We have developed novel computational algorithm that leverages Rice's formula to construct an importance biasing distribution that helps us compute the extreme excursion probabilities efficiently.

Joint work with Vishwas Rao, Argonne National Laboratory.

29. SimILE (Simulated Implicit Likelihood Estimation): Bayesian Analyses Using Estimated Implicit Likelihoods with Discretized Simulated Data

Michael Hamada, Los Alamos National Laboratory

We present a Bayesian inferential method where the likelihood for a model is unknown (i.e., implicit) but where data can easily be simulated from the model. We discretize simulated (continuous) data to estimate the implicit likelihood in a Bayesian analysis employing a Markov chain Monte Carlo algorithm. Some examples are presented as well as a small study on some of the method's properties.

30. Data Fusion, Modelling, and the Spread of Modern Humans

Rachel J. A. Hopkins, University of Oxford (UK) & University of New Mexico (USA)

The spread of Anatomically Modern Humans (AMHs) and the coinciding demise of the Neanderthals in Europe (50-40k years ago) has fascinated researchers since the dawn of archaeology and the investigation of human evolution. However, archaeological datasets are inherently messy and sampling is affected by 'survival bias'. We will therefore demonstrate how data fusion, innovative Bayesian modelling, and Kernel Density Estimation (KDE) can drastically improve the quality of data interpretation by integrating qualitative and quantitative evidence. The resulting spatio-temporal models indicate that AMHs migrated into central Europe 5-7k years earlier than previously estimated. These findings change our understanding of the conditions under which Neanderthals and AMHs may have interacted, and challenge the often-perpetuated dichotomy in their material culture. This work is simultaneously innovating on diagnostic and analyses techniques to constrain interpretation on multiple fronts. We applied recent improvements in radiocarbon dating methodologies (e.g. ultrafiltration and single amino acid dating) and designed targeted sampling strategies, to reduce noise deriving from exogenous carbon contamination and non-human activities at archaeological sites. For analysis, we adapted Bayesian modelling and KDE techniques to integrate different qualitative and quantitative data from geophysics, forensics, physical anthropology, genetics, palaeoproteomics and material culture -- work that included the development of new analyses and visualisation tools, such as KDE mapping of geo-located radiocarbon information. Our research generated pioneering high-resolution chronologies for major archaeological sites along the Danube fluvial corridor, and the first pan-European spatio-temporal mapping of AMH and Neanderthal activity that includes reliable data from not only western and central European, but also eastern European sites. As a result, we increased the reliability of the AMH dispersal chronology, as well as refined our understanding of changing migration patterns over large geographical areas, and of the regional variability and complexity of the period between 50-30k BP (Before Present).

Joint work with Bibaná Hromadová (CNRS, Nanterre, France), András Markó (Hungarian National Museum, Budapest), Jean-Luc Guadelli (CNRS, Bordeaux, France), Nikolay Sirakov (National Institute of Archaeology and Museum, Bulgarian Academy of Sciences), Aleta Guadelli (CNRS, Bordeaux, France), Christopher Bronk Ramsey (University of Oxford, UK), Tom F. Higham (University of Oxford, UK).
Acknowledgments: This research was funded by the European Research Council under the European Union's 7th Framework Programme (FP7/2007-2013) / ERC grant agreement no. [324139] "PalaeoChron" awarded to Prof Tom Higham; the National Institute of Archaeology of the Bulgarian Academy of Sciences; and UMR5199 CNRS PACEA (France). Wolfson College Oxford provided travel contributions, and the Hunt Fellowship (Wenner Gren Foundation) supports the publication process. Access to material for radiocarbon dating was granted by: Dr Walpurga Antl-Weiser, Prof Dr Maria Teschler Nicola, Dr Christine Neugebauer-Maresch, Dr L'ubomíra Kamínska, Dr Matej Ruttkay, Dr Alena Šefčáková, Dr Anna Durisova, Dr Gasparik Mihaly, Dr Klára Paloás, Dr László Makádi, Dr Dušan Borić, and Dr Emanuela Cristiani.

31. Penalized Ensemble Kalman Filters for High Dimensional Non-linear Systems

Elizabeth Hou, University of Michigan

The ensemble Kalman filter (EnKF) is a data assimilation technique that uses an ensemble of models, updated with data, to track the time evolution of a usually non-linear system. However, its performance can suffer when the ensemble size is smaller than the state space, as is often necessary for computationally burdensome model. To solve this problem in this high dimensional regime, we propose a computationally fast and easy to implement algorithm called the penalized ensemble Kalman filter (PEnKF). Under certain conditions, it can be theoretically proven that the PEnKF will have good performance (the error will converge to zero) despite having fewer ensemble members than state dimensions. Further, the proposed approach makes fewer assumptions about the structure of the system than localization methods. These theoretical results are supported with simulations of several non-linear and high dimensional systems.

Joint work with Earl Lawrence, Alfred Hero.

32. Characterizing Dislocation Defects in Dark Field X-ray Microscopy Images of Bulk Diamond

Marylesa Howard, Nevada National Security Site

Material defects play a large role in material response under shock loading, yet our understanding of how these defects initiate, propagate, and annihilate are not well understood. Using a newly developed diagnostic, dark field X-ray microscopy (DFXM), we can now visualize the behavior of dislocation defects in materials at the mesoscale under varying conditions. Using data from the European Synchrotron Radiation Facility, we apply a variety of image processing techniques to capture relevant features to locate and characterize size and orientation of dislocation defects in bulk diamond DFXM images. Beyond simple visualization, this analysis drives statistical characterization of the defects and their dynamic response to further improve the relevant physics models.

This work was done by Mission Support and Test Services, LLC, under Contract No. DE-NA0003624 with the U.S. Department of Energy and supported by the Site-Directed Research and Development Program. DOE/NV/03624--0681

Joint work with Ryan Coffee, SLAC Malena Español, Arizona State University Arnulfo Gonzalez, NNSS Jessica Pillow, NNSS/University of Arizona Michael Brennan, MIT Eric Machorro, PNNL Eric Clarkson, University of Arizona Sean Breckling, NNSS Margaret Lund, NNSS Jesse Adams, NNSS Leora Dresselhaus-Cooper, LLNL.

33. Scalable hyperparameter tuning for machine-learning emulation of a seismic Earth model
Derek Jensen, Lawrence Livermore National Laboratory

Three-dimensional Earth models are able to estimate seismic-phase travel times significantly more accurately than 1D models, achieving very low global travel-time prediction errors which allows for more accurate estimates of event epicenters. Despite this advantage, 3D models can be too computationally expensive for use in real-time analysis. The objective of the current work is to robustly emulate seismic-phase travel times computed using a 3D Earth model with computational speeds that are fast enough for real-time analysis and with prediction errors that that are insignificant compared to the inherent uncertainty of the Earth model. To this end, a deep neural network (DNN) was trained and tuned on millions of realizations of a 3D Earth model using the Ray Tune framework to efficiently construct a DNN architecture. The best performing DNNs achieved acceptable errors, although we observed significantly larger errors associated with short-path rays and rays that pass near the core of the Earth. Additional data are begin generated in an effort to improve predictions in these distance ranges. We plan to release the final model as a TensorFlow model.

Joint work with Stephen Myers, Clifford Anderson-Bergman. Prepared by LLNL under Contract DE-AC52-07NA27344 and released under IM number: LLNL-ABS-803559

34. Using Local Activity Models to Improve Object Detection
Stephen T. Jones, Sandia National Laboratories

Recent advances in machine learning have produced great strides in the performance of image-based object detectors. Nevertheless, for a variety of reasons ranging from challenging backgrounds to limitations on training examples for unusual objects, performance limitations remain. In this work, we provide an initial demonstration that local activity models can improve the overall performance of object detectors by predicting the expected locations of target objects based on historical observations. We use synthetic data to demonstrate the efficacy of our approach and discuss directions for future research.

Joint work with David Stracuzzi, Sandia National Laboratories Charlie Vollmer, Sandia National Laboratories.

35. Spatio-temporal analysis of endemic histoplasmosis
Kimberly Kaufeld, Los Alamos National Laboratory

Histoplasmosis is an infection caused by inhaling fungal spores that typically dwell in soil containing large amounts of bat or bird droppings. In healthy individuals, infection typically leads to a mild fever and cough, but in immunocompromised individuals, symptoms can be severe and even deadly. Histoplasmosis is the most common fungus endemic to North America, and the number of cases has been on the rise over the past decade with the endemic region appearing to expand as environmental conditions change. However, surveillance is challenging as few states report to the Centers for Disease Control and Prevention (CDC), and those that do can only report the cases that present for care. Using the monthly CDC data from 2010-2014 for the states reporting, we estimate changes in the endemic region of histoplasmosis. We fit a spatio-temporal occupancy model to allow us to model both the probability that a case was detected and the probability that histoplasmosis is present in each county. We incorporate information about environmental factors to better characterize areas where histoplasmosis is likely to be endemic. This work aims to help inform physicians to screen for histoplasmosis, particularly in newly endemic regions.

Joint work with Staci Hepler, Wake Forest University David Kline, The Ohio State University.

36. Is There ____ on Mars?: Disaggregation of LIBS Spectra
Natalie Klein, Los Alamos National Laboratory

As part of its mission, the Mars rover Curiosity is equipped with ChemCam, an instrument designed to determine the composition of soil and rocks using laser-induced breakdown spectroscopy (LIBS). Traditional approaches to disaggregation, or quantification of component materials within a sample, are based on regression techniques trained on a library of LIBS spectra with known material compositions. However, because samples are expensive to create, the number of observations tends to be small and the design space is focused around particular types of materials. To explore a larger space of materials, we leverage a physics simulation code that can predict spectra given the plasma temperature and density and the proportions of constituent materials. Using both supervised machine learning techniques and Bayesian model calibration, we show that fusion of simulation outputs with real data is beneficial for disaggregation. In addition, our models quantify uncertainty and address systematic discrepancies between simulation outputs and real data, leading to better understanding of mismatches between theory and data.

37. Analyzing similarity patterns in seismic data
Goran Konjevod, Lawrence Livermore National Laboratory

We present methods to find and analyze patterns in seismic data series. As a preprocessor, we use FAST [Bailis et al], developed as an unsupervised earthquake detection pipeline. However, since our interest is primarily in detecting signals and patterns that are generated by human activity, we need to go beyond the similarity search and detection methods of FAST.

We use the output of FAST to create a weighted similarity graph (or in the case of data collected from multiple sensors, a collection of graphs), where nodes correspond to short segments of time. Then an edge between two nodes expresses the similarity between the signal behavior during the time segments corresponding to the two nodes. We analyze substructures arising in such graphs, helped somewhat by the fact that the temporal ordering of nodes is known. The additional structure allows us to compare sequences of nodes instead of just individual ones, and to address the problem of searching for structured and recurrent activity patterns.

We also consider the problem of fusing the data collected by separate sensors whose responses may differ significantly even though they may be located in close vicinity of one another.

Joint work with Gabriel Andrade, CU Boulder Jose Cadena, LLNL.

38. Data Collection and Analysis Challenges and Mitigation Strategies for Quantitative Human Factors Research Studies in Nuclear Power Plant Modernization
Casey Kovesdi, Idaho National Laboratory

The United States (U.S.) Department of Energy (DOE) Light Water Reactor Sustainability (LWRS) program Plant Modernization Pathway is conducting targeted research and development (R&D) to address aging and reliability concerns with the legacy instrumentation, control ,and related information systems of the U.S. LWR fleet. In this effort, the application of human factors engineering (HFE) provides an important role in ensuring new digital plant technologies enable broad innovation and business improvement with continued operational safety. Evaluation is a key activity in HFE, which often occurs iteratively through the system design lifecycle. While qualitative methods are important to characterize users' experiences, attitudes, behaviors and interactions, it is important to complement qualitative approaches with objective quantitative data collected in controlled experiments. In collecting and analyzing this quantitative data, there are notable challenges in the nuclear HFE domain that make it difficult to use standard methods and techniques to make valid and reliable inferences. Notable challenges include small sample size, limited resources, large error variance, heterogeneity of variance, small effect sizes, ceiling effects, and multivariate mixed type response distributions. The challenges often result in control room studies with statistically non-significant results, limiting quantitative inferences. This work discusses these challenges across different scientific viewpoints and provides real-world examples of these challenges in practice. Collectively, the objective of this work is to position these challenges to the larger data science community as a means of identifying future opportunities to address these issues.

Joint work with {Katya Le Blanc, Andrea Mack, & Ruixuan Li} - Idaho National Laboratory.

39. Sensitivity analysis for Monte Carlo simulations of integro-partial differential equations
Rich Lehoucq, Sandia National Laboratories

Monte Carlo methods for integro-partial differential equations sample the underlying stochastic process generating the equation. As the sample size increases, improved approximations of the integro-partial differential equations result. Monte Carlo methods are typically used for equations with more than three dimensions or when the quantity of interest only involves a portion of the solution.

My poster will review the problem of computing derivatives for the various quantities of interest estimated during the Monte Carlo simulations. These derivatives, or sensitivities, pave the path towards a currently unavailable capability of design, optimization and uncertainty quantification, staples of modeling and simulation when only Monte Carlo simulation is available.

40. Statistical Modeling for Spatio-Temporal Data from Physical Convection-Diffusion Processes
Xiao Liu, Department of Industrial Engineering, University of Arkansas

This paper proposes a statistical modeling approach for spatio-temporal data arising from non-linear convection-diffusion processes. Such processes are widely found in scientific and engineering applications where governing physics imposes critical constraints on how data can be modeled and how statistical models should be interpreted. The idea of spectrum decomposition is employed to approximate a physical process by the linear combination of spatial basis functions and a multivariate stochastic process of spectral coefficients. We consider spatially-varying convection-diffusion and source-sink. As a result, the temporal evolution of spectrum coefficients is coupled with each other, which corresponds to the non-linear transfer or redistribution of energy across multiple scales from the perspective of physics. Because of the spatially-varying convection-diffusion, the space-time covariance is non-stationary in space, while the existing PDE-based statistical models still primarily focus on stationary processes. The theoretical results are integrated into the framework of hierarchical dynamical spatio-temporal models. Some considerations related to computational efficiency are discussed in order to make the proposed approach practical for real-time operations. The advantages of the proposed methodology are demonstrated by a case study and comparison studies.

Joint work with Kyongmin Yeo, IBM T. J. Watson Research Center; Siyuan Lu, IBM T. J. Watson Research Center;.

41. Understanding the Merits of Winning Solutions from a Data Competition for Varied Sets of Multiple Criteria
Lu Lu, University of South Florida

In the Topcoder data competiton focused on urban radiation search (https://www.topcoder.com/lp/detect-radiation) recently hosted by Los Alamos National Laboratory, the goal of the competition was to identify top solutions to successfully detect, identify and locate one of six radioactive materials located on a U.S. urban street. The top competitors used different approaches with varied strengths and weaknesses. If one of these solutions was to be selected for deployment in practice, the details of the situation would be highly relevant to match the best available solution to the relative emphasis of which material was thought to be the highest priority to find, as well as how important each of objectives for detection, identification and location were. In this poster, we examine the robustness of the performance of competitors' solutions across different scenarios and highlighted priorities. Using Pareto fronts, it is possible to eliminate some solutions that are never best for the defined priorities, as well as choose best solutions for a subset of the possible objectives, and also gain understanding about the relative merits of the top solutions.

Joint work with Dr. Christine M. Anderson-Cook from Statistical Sciences Group at LANL Miaolu Zhao from University of South Florida.

42. Condition-based Maintenance of Nuclear Power Plant Secondary Piping Structures with Machine Learning

Andrea Mack, Idaho National Laboratory

The current aging management plans of passive structures in nuclear power plants (NPPs) are based on preventative maintenance strategies. These strategies involve periodic, manual inspection of passive structures using nondestructive examination techniques. This manual approach is prone to errors. It also contributes to high operation and maintenance costs, making it is cost prohibitive. To address these concerns, a transition from the current preventive maintenance strategy to a condition-based maintenance strategy is needed. The research presented in this paper develops a condition-based maintenance capability to detect corrosion in secondary piping structures in NPPs. To achieve this, a data-driven methodology is developed and validated for detecting a surrogate corrosion processes in piping structures. A scaled-down experimental test bed is developed to evaluate the corrosion process in secondary piping in NPPs. The experimental test bed is instrumented with tri-axial vibration accelerometers and data is collected under different operating conditions. Data are processed using the Hilbert-Huang Transformation. Distributional features of phase information among the accelerometers were used as features in support vector machine (SVM) and least absolute shrinkage and selection operator (LASSO) logistic regression methodologies to detect changes in the pipe condition from its baseline state. SVM classification accuracy averaged 99% for all models. LASSO classification accuracy averaged 99% for all models using X-direction data.

Joint work with Koushik Araseethota Manjunatha and Vivek Agarwal (Idaho National Laboratory); David Koester and Douglas Adams (Vanderbilt University).

[1] Kansara, D. P., Sorathiya, A. P., and Patel, H. R., 2018. 'Corrosion monitoring and detection techniques in petrochemical refineries'. IOSR: Journal of Electrical and Electronics Engineerings,13-2.

[2] Aalsalem, M. Y., Khan, W. Z., Gharibi, W., Khan, M. K., and Arshad, Q., 2018. 'Wireless sensor networks in the oil and gas industry: recent advances, taxonomy, requirements, and open challenges'. Journal of Network and Computer Applications,113, pp. 87-97.

[3] Park, G.Y., Lee, C.K., Kim, J.T., and Kwon, K.C., 2006. 'Pipe corrosion analysis by time-frequency distribution and ridge pattern'. pp. 1570-1573.

[4] Munavalli, S. C., Pissinou, N., Lagos, L. E., and Jin, X., 2013. 'Structural damage detection of nuclear reactor sites using sensor networks'. pp. 1-4.

[5] Hing, Y., Garrett Jr., J.H., Oppenheim, I.J., & Soibelman, L., 2013. 'Toward data-driven structural health monitoring: application of machine learning and signal processing to damage detection'. Journal of Computing in Civil Engineering, 27(6), November.

[6] Salehi, H., and Burgueno, R., 2018. 'Emerging artificial intelligence methods in structural engineering'. Engineering Structures,171, pp. 170-189.

[7] Huang, N., and Shen, S., 2014. Hilbert-Huang Transform and Its Applications, Vol. 16. World Scientific.

[8] Suykens, J. A., and Vandewalle, J., 1999. 'Least squares support vector machine classifiers'. Neural processing letters,9(3), pp. 293-300.

[9] Hastie, T., Tibshirani, R., and Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

43. Heterogeneous Data Streams for Predicting Mosquito Density

Carrie Manore, Los Alamos National Laboratory

While for many diseases incidence is decreasing, human disease cases from mosquito, tick and flea bites have tripled in the US in the past decade. A For mosquito-borne diseases such as West Nile virus (WNV) and Eastern equine encephalitis (EEE), effective mosquito control is our main mitigation option. However, understanding when are where mosquito populations are increasing is key to effective mosquito control. Identifying reliable approaches to predicting fluctuations in mosquito population size has been challenging due to the interdependence of mosquito development and environmental conditions. We take a data fusion approach to bring together laboratory observed drivers of mosquito development and field measurements of daily water gauge levels, average temperature, and average maximum daylight hours to inform our process based model. We accurately predict 5 years of test data in the Greater Toronto Area.

44. Inferring Influence Networks from Longitudinal Bipartite Relational Data

Frank W Marrs, Los Alamos National Laboratory

Longitudinal bipartite relational data characterize the evolution of relations between pairs of actors, where actors are of two distinct types and relations exist only between disparate types. A common goal is to understand the temporal dependencies, specifically which actor relations incite later actor relations. There are two primary existing approaches to this problem. The first projects the bipartite data in each time period to a unipartite network and uses existing unipartite network models. Unfortunately, information is lost in calculating the projection and generative models for networks obtained through this process are scarce. The second approach represents dependencies using two unipartite influence networks, corresponding to the two actor types. Existing models taking this approach are bilinear in the influence networks, creating challenges in computation and interpretation. We propose a novel generative model that permits estimation of weighted, directed influence networks and does not suffer from these shortcomings. The proposed model is linear in the influence networks, permitting inference using off-the-shelf software tools. We demonstrate the performance of the proposed model in simulation studies and an analysis of weekly international state interactions.

Joint work with Benjamin W. Campbell, Bailey K. Fosdick, Skyler J. Cranmer, Tobias Böhmelt.

45. Exploring the Impact of Different Scaling of Weights for Non-Uniform Space Filling Designs

Miranda Martin, Los Alamos National Laboratory

Traditionally, space filling designs have sought to provide uniform density of points throughout the region of interest. The Non-Uniform Space Filling (NUSF) capability introduced in Lu, Anderson-Cook, Ahmed (2019), allows the experimenter to change this fundamental assumption and to increase the proportion of design points in regions close to the anticipated optimum location, near interesting features, or where increased precision in prediction is sought. In this poster, we describe two alternate approaches for defining the desired density of points throughout the input space. By providing weights based on a direct measure of desirability or an alternative based on ranking, the experimenter now has greater control of design construction. We describe the NUSF approach, illustrate differences between the two weighting methods, and provide guidance for how to assess which method is better suited to the goals of the experiment.

Joint work with Towfiq Ahmed, LANL Christine M. Anderson-Cook, LANL Lu Lu, University of South Florida.

46. Automated algorithms for predicting trends and identifying subpopulations in neutron generator (NG) production data

R. Multari, Ph.D., Sandia National Laboratories

Results from current research to create automated algorithms for the rapid, real time analysis of neutron generator production data will be presented. The analysis methods used in this investigation focus on methods effective at extracting signal from the data even when the signal is not visually discernable. Multivariate analysis techniques and wavelet analysis methods will be used to iteratively develop predictive models to differentiate normal neutron generator production data from abnormal production data and to identify the type of abnormality detected. Such a capability would significantly reduce the engineering resourcing required to review neutron generator manufacturing data and would enable the early detection of trending manufacturing data. The results of this research can be applied to all levels and types of neutron generator manufacturing and may also be useful for improving other commercial manufacturing processes.

This research was inspired by the successes in the field of spectroscopy in utilizing multivariate and statistical analysis methods to build predictive algorithms capable of correctly segregating input samples into groupings even when there was no visible difference between spectra and in the presence of a variable spectral background.

In this work, the multivariate analysis techniques under investigation focus on predicting performance using all manufacturing data arranged to create a ‘fingerprint’ of the unit under test. Fingerprints are then grouped according to the performance to be detected and predictive models for each group are created using principal component partial least square regression to predict which group new input data belongs to. The wavelet analysis techniques under investigation focus on decomposing and localizing patterns in the data and constructing classifiers to discriminate between data set groups. Time series data is transformed into a family of wavelets and the coefficients from this decomposition are used to discriminate between the data set groupings. The models form these investigations with best predictive performance on previously identified test data and new production data can be combined in a logical, programmed flow to create an automated algorithm to screen for abnormal and trending NG manufacturing data.

It should be again noted that the algorithm development methodology created in the course of this project can also be applied to other manufacturing processes to create similar capability.

Joint work with J. Ray, L. Miller, R. Ferrizz, P. Cummings, C. Co, L. Walla, and S. Martin.

47. Non-negative Tensor Factorization for Isolation of Seismic Signals
Ben Nebgen, Los Alamos National Laboratory

Seismic sensors are able to observe signals from many different sources, such as earth quakes, construction, wind farms, and many others. The ability to isolate the frequency spectrum produced by one of these sources from the noisy datastream is a difficult problem. In an effort to automate this process, we apply the method of non-negative tensor factorization (NTF) to preprocessed seismic data. NTF is an unsupervised Machine Learning (ML) technique that decomposes a tensor into latent features that are obfuscated in the original dataset. The non-negativity of NTF approaches results in highly interpretable features, such that experts are easily able to identify the sources of the latent features. We present the results of NTF applied to seismic data obtained from Northern Texas, from which multiple anthropogenic noise sources are isolated.

Joint work with Elizabeth Mallory, Jon MacCarthy, Omar Marcillo, Kim Rasmussen, Boian Alexandrov.

48. FAST : Finding Anomalous Ship Trajectories
Benjamin Newton, Sandia National Laboratories

Huge amounts of data are collected daily capturing the paths of ships moving through oceans and waterways throughout the world. Most of these ship trajectories (or tracks) are fairly boring, with ships often taking essentially straight-line paths between ports, with occasional exceptions when required to maneuver through or around some obstacle. Using Tracktable, a tool developed by Sandia National Laboratories to analyze trajectories, we decompose each trajectory into a set of features, and apply machine learning techniques to these features to quickly find anomalous ship trajectories.

Joint work with M. Daniel Rintoul (Sandia National Laboratories), Andrew Wilson (Sandia National Laboratories), Robert Warrick (Sandia National Laboratories), Jessica Jones (Sandia National Laboratories).

49. Causal inference modeling for feature engineering of QSAR machine learning models
Bernard Nguyen, Sandia National Laboratories

Molecular descriptors are commonly used to digitally represent the physical structure of a molecule in quantitative structure-activity relationship (QSAR) machine learning models, but the overwhelming abundance of features can negatively impact the model's predictive performance. This research explores causal inference modeling as a method of feature engineering for ensemble-based machine learning techniques, and evaluates model performance against other common methods of feature evaluation.

50. Interleaved Analytics and Simulation: An Expanded View of In Situ Computing
George Ostrouchov, Oak Ridge National Laboratory

Simulation science codes on leadership computing systems, like the Oak Ridge National Laboratory's Summit, often produce more data than is analyzed within a given scientific campaign. User facilities, like the Spallation Neutron Source, also continually increase their data production. Analytics and learning are often treated as a distinct phase of investigation either succeeding or preceding data generation. Traditional approaches focus on data decimation and data compression 'in situ' to make I/O easier and enable visualization and simulation steering. We take an expanded view of interleaving analytics, simulation, and possibly external data. Various data analysis, machine learning (ML), and deep learning paradigms in HPC bring new challenges to in situ computing. The challenges include automation, since data analysis is typically a discovery process needing user interaction and iteration, and reconfiguration of the analysis itself to accommodate streaming data. The workflows required to orchestrate each situation can seem disparate and even unique. We investigate the common aspects of these workflows that can drive the development of useful infrastructure for interleaved data analysis and simulation on large systems.

Joint work with Drew Schmidt, Michael Matheson, Junqi Yin, Aristeidis Tsaris, Benjamin Hernandez, and Mallikarjun Shankar Oak Ridge National Laboratory.

51. Development and Demonstration of Data-driven Compact Device Models for Circuit Simulation and Analysis
Biliana S. Paskaleva, Sandia National Laboratories

Compact semiconductor device models are essential for efficiently designing and analyzing large circuits. However, traditional compact model development requires a large amount of manual effort and can span many years. Moreover, inclusion of new physics (e.g., radiation effects) into an existing compact model is not trivial and may require redevelopment from scratch. Machine Learning (ML) techniques have the potential to automate and significantly speed up the development of compact models. In addition, ML provides a range of modeling options that can be used to develop hierarchies of compact models tailored to specific circuit design stages.

In this work, we explore three such options: (1) table-based interpolation (TBI), (2) Generalized Moving Least-Squares (GMLS), and (3) feed-forward Deep Neural Networks (DNN), to develop compact models for a p-n junction devices. Specifically, we apply these approaches to develop compact models for common mass produced semiconductor devices with well-documented electrical and thermal characteristics.

The first approach, TBI, is a local parametric regression technique that uses cubic splines to construct a piecewise polynomial approximation of available electrical data. TBI is used extensively in many modeling contexts, offering simplicity, computational efficiency, and the ability to generate differentiable approximations. Drawbacks include significant memory requirements and datasets restricted to rectangular grids. The second approach uses GMLS approximants to build compact device models; this method, unlike TBI, can be applied to scattered data as well. GMLS is an example of non-parametric regression, which uses local kernels to build estimates from scattered data. We believe that we are the first to apply GMLS to compact device modeling and demonstrate them in circuit simulations.

Finally, we develop DNN device models. The success of DNNs in various classification tasks is well documented; however, their application to scientific computing and modeling is more recent, generating growing interest. Our work provides important insights into the development of DNN compact models, informed by performing circuit-level simulations using such data-driven devices. A key contribution of our work is the development of a DNN training strategy, based on transformed sets of electrical measurements, that consistently produce physically correct compact diode models across a range of DNN architectures.

We evaluate the performance of these ‘data-driven’ compact models by (1) comparing their voltage-current characteristics against laboratory data, and (2) building simple circuit using these devices, predicting the circuit’s behavior using SPICE-like circuit simulations, and then comparing these predictions against laboratory measurements of the same circuit.

Joint work with K. Aadithya, P. Kuberry, P. Bochev, K. Leeson, A. Mar, T. Mei, and E. Keiter,

52. Molecule Generation with Deep Reinforcement Learning for Geometric Structures at Exascale
Jenna Pope, Pacific Northwest National Laboratory

Many important problems in biology, chemistry, and materials science can be solved through the design of novel molecular structures with tailored properties. However, the automated generation of novel and valid molecules that optimize towards the desired properties is a challenging task. The challenge stems from the combinatorial explosion that rises from the sheer number of possible structures that needs to be explored. The multi-laboratory ExaLearn project, part of the DOE Exascale Computing Project, aims to produce a novel 'Design' capability that combines state-of-the-art machine-learning techniques with high-performance computing to enable the rapid exploration of chemical space on Exascale-class systems. Humans are naturally trained to think of chemical structures as a collection of atoms and molecules connected via different bonds, and graph theory provides a natural tool for capturing that intuition in a concrete mathematical fashion. However, most successful deep learning techniques such as convolutional neural networks were designed for grid-based, Euclidean data sources such as images and video data, but atoms in molecules are not restricted to a grid. Therefore, automated design of molecular graphs requires exploration of novel graph-based machine-learning approaches that satisfies unique modeling needs for molecular structures. This poster will present early results from a scalable deep reinforcement learning workflow aimed at learning novel geometric structures. ExaDesign implements two approaches for deep reinforcement learning (RL) on attributed graphs. The first one is based on the MolDQN framework that uses deep Q-learning and randomized value functions. The second approach aims to integrate graph neural network and deep Q-learning as the policy network. Additionally, we are developing a number of graph-theoretic reward functions that characterize target molecular properties for both these modeling approaches. Graph theoretic chemical descriptors are a very well-established concept in the computational chemistry research community, and we believe their use to steer the output from RL algorithms will lead to more interpretable results. Both RL models are trained via a surrogate model for the domain. The surrogate model is trained using a message passing graph neural network. The RL models are trained in a distributed fashion by the ExaRL framework developed by the DOE ExaLearn program. Our poster presentation will showcase results from two distinct applications in chemistry. The first one focuses on quantitative modeling of intermolecular interactions within liquid water. Reasoning about how water molecules interact at large scale going through phase transitions is critical for gaining a full understanding of chemistry in aqueous environments. Our second application focuses on designing molecules for performant and safe electrolytes in next-generation Li-ion batteries.

Joint work with Logan Ward (Argonne National Laboratory), Marcus Schwarting (Argonne National Laboratory), Joseph Heindel (University of Washington), Sotiris Xantheas (Pacific Northwest National Laboratory and University of Washington), Sutanay Choudhury (Pacific Northwest National Laboratory), Malachi Schram (Pacific Northwest National Laboratory).

53. Robust Multi-modal Event Detection by Weak Matching of Temporal Interval Graphs of Observables
Lakshman Prasad, Los Alamos National Laboratory

A complex event, or process, is typically characterized by disparate observables manifesting at different times, over different durations, and with temporal patterns of occurrences over the event's duration. Our earlier work introduced an interval graph representation framework for multi-modal events. We represented events or processes as temporally evolving interval graphs, where event observables are captured as attributed time intervals represented as nodes, and temporally proximate observables' nodes are joined by edges. Comparing events can then be posed as a graph matching problem. However, graph matching is a notoriously difficult problem that is NP-hard in general, with polynomial algorithms only for very restricted families of graphs. Even in these cases, perturbations to graph structure from missing or extra nodes and edges can lead to poor matching results. Indeed, real-world sensing often involves noisy environments where extraneous observables or missed ones may interfere with the event graph structure. Motivated by this need for robust detection of temporal signatures of real-world events in the presence of irrelevant observables and in the absence of relevant observables, we present here a noise-tolerant, weak-matching algorithm for interval graphs in the presence of topological noise. We employ a weak representation of event interval graphs via their shortest and longest paths and compare them by matching their path sets. The paths in the graphs can be interpreted as event signatures. We describe a node and edge-attributed path matching scheme that is robust to inclusions and exclusions of nodes by adapting the longest common subsequence algorithm using dynamic programming for attributed path matching. Our graph representation and matching scheme is targeted at dynamically detecting activities such as nuclear proliferation, terrorism, and epidemics, using sensor and social media information in conjunction with subject matter expertise and event/process models.

Prasad .L, 'Multiscale graph-based framework for efficient multisensor integration and event detection,' Proc. SPIE 10646, Signal Processing, Sensor/Information Fusion, and Target Recognition XXVII, 106460Q (27 April 2018); doi: 10.1117/12.2304288,SPIE Defense + Security, 2018, Orlando, Florida, United States

54. Sensor Measurement Error Estimation for Reactor Coolant Systems Using Information Fusion Methods
Pradeep Ramuhalli, Oak Ridge National Laboratory

State estimation of a nuclear power plant is important for supporting decisions on actions for a safe continued operation. Extensive instrumentation is used to measure critical variables such as temperature, pressure, flow rate, and radiation levels. The accuracy of these measurements is critical for quantifying the health of the plant and its components. Over time, aging sensors can result in slow drifts in their calibration which lead to gradual errors that are often too subtle to detect by simple threshold methods. In addition, measurement errors can be due to factors such as blockage, leaks, and air intrusion in sensing lines, process and electronic noise, and electromagnetic interference. These errors, if not identified and addressed, can negatively impact the decisions on critical operational actions. We consider a simplified version of a sensor error estimation problem for the primary coolant loop of a pressurized water reactor. We present an information fusion method to estimate the sensor errors by combining measurements from multiple sensors, and evaluate its performance for pressure sensors of the coolant loop. The sensors measure the plant variables that are related to each other by physical laws, which are based on the fluid flow and thermodynamic properties of the primary coolant loop. By exploiting these relationships, we estimate a plant variable by fusing measurements from sensors that measure the same or other variables. In particular, the underlying relationship, between sensor measurements and the selected variable, is learned as a regression function using measurements collected under controlled plant conditions. We present the ensemble of trees and support vector machine fusers which are trained using sensor measurements collected over an emulated test loop of a pressurized water reactor. We assess the error estimates for pressure transmitters at the heat exchanger of the primary coolant system under twenty controlled scenarios. Positive and negative errors are captured by both methods under externally introduced calibration drifts, blockages, minor leaks and air gaps in sensing lines, and electromagnetic interference; the root mean square error is typically within 2% percent of the maximum measurement. To complement the experimental results, we present analytical results that characterize the generalization performance of the sensor error estimates on future measurements under the distribution-free machine learning formulation. We derive the generalization equations that establish that the underlying problem is effectively solvable by the two methods, and their form reflects the specific estimator properties. By utilizing smoothness (bounded Lipschitz constant) and non-smoothness (fixed number of finite jumps) properties, they guarantee that the generalization error is bounded by a specified parameter with a confidence probability that improves with the number of measurements.

Joint work with Nageswara S. V. Rao, Christopher Greulich, Sacit M. Cetiner.

55. Analytics of Wide-Area MPI Measurements

Nagi Rao, Oak Ridge National Laboratory

Scientific workflows are increasingly being distributed across wide-area networks, and their code executions are expected to span across geographically dispersed computing systems, including clusters and supercomputers. Message Passing Interface (MPI) has been extensively used to support communications for distributed computations, typically, over compute clusters and high-performance systems within a single facility. We study the performance of MPI basic operations over long distance connections, wherein Transmission Control Protocol (TCP) is used for the underlying network transport. We describe measurements of execution times of MPI codes that utilize MPI Sendrecv operations over emulated 10Gbps connections with 0-366ms Round Trip Time (RTT), including the longest one spanning the globe. We describe measurements under periodic and random losses with uniform, Poisson and Gaussian distributions. They demonstrate that basic MPI codes can be sustained over long distance connections under external packet loss rates up to 20%. These measurements highlight the qualitative effects of losses which manifest as increased execution times as a consequence of TCP's loss recovery process. Indeed, application-level measurements of MPI operations reflect the network connection RTT and loss rate, which should be accounted for effective distributed computations, for example, by avoiding long, highly congested connections.

We study five disparate machine learning methods, with linear and non-linear, and smooth and non-smooth properties, to estimate the connection RTT and loss rate using execution time measurements of MPI Sendrecv operations. They include four non-linear estimators, namely, smooth support vector machine and Gaussian Process Regression (GPR), and non-smooth ensemble of trees and regression trees, in addition to the baseline Linear Regression (LR) method. The diversity and complexity of these estimators combined with the randomness of losses and TCP's non-linear response rule out the identification of a single best among the estimators. Instead, we fuse them to retain their design diversity using Linear Regression fusion (LR-F) and GPR fusion (GPR-F) methods. Among individual RTT estimates, GPR has the lowest Root Mean Square Error (RMSE) in all four loss scenarios, and GPR-F fuser achieved even lower RMSE in three out of four loss scenarios while encompassing the design diversity of the individual methods. Overall, our results show that accurate estimates can be generated at low loss rates but become inaccurate at loss rates of 10% and higher, wherein the datasets appear much too complex for these methods. In addition, they reveal some subtle performance effects including over-smoothing by some estimators in achieving lower RMSE, and bleeding effects of RTT in loss rate estimates. We develop analytical characterization of the performance improvements of fused estimates over individual RTT estimates under finite sample, dis

Joint work with Neena Imam, Oak Ridge National Laboratory; Zhenchun Liu, Raj Kettimuthu, Ian Foster, Argonne National Laboratory.

[1] N. S. V. Rao, N. Imam, Z. Liu, R. Kettimuthu, I. Foster, Machine learning methods for connection RTT and loss rate estimation using MPI measurements under random losses, International Conference on Machine Learning for Networking (MLN'2019), Paris, France, December 3-5, 2019.

[2] N. S. V. Rao, N. Imam, Z. Liu, R. Kettimuthu, I. Foster, Estimation of RTT and loss rate of wide-area connections using MPI measurements, IEEE/ACM Workshop on Innovating the Network for Data-Intensive Science (INDIS2019), Denver, CO, November 17, 2019.

[3] N. S. V. Rao, N. Imam, S. Boehm, A case study of MPI over long distance connections, 13th Annual IEEE International Systems Conference (SYSCON2019), April 8-11, 2019, Orlando, Florida.

56. Assessing the reliability of prediction intervals from Bayesian Neural Networks

Daniel Ries, Sandia National Laboratories

Neural networks (NN) have become popular models because of their predictive power in a variety of applications. Users are beginning to use NN to automate tasks previously done by humans. One criticism of NN is they provide no uncertainty with their predictions, which is problematic in high risk applications. Bayesian neural networks (BNN) provide one approach to quantifying uncertainty by putting NN in a probabilistic framework through placing priors on all weights and computing posterior predictive distributions. We assess the quality of uncertainty given by BNN estimated using Markov Chain Monte Carlo (MCMC) and variational inference (VI) with a simulation study. These results are also compared to Concrete Dropout, another way to provide uncertainty for NN, and to a Gaussian Process model. The effect of network architecture on uncertainty quantification is also explored. BNN fit via MCMC gave uncertainty results similar to those of the Gaussian Process, which performed better than BNN fit via VI or Concrete Dropout. Results also show the significant effects of network architecture on interpolation and show additional issues with over- and underfitting.

Joint work with Philip Lewis, Josh Zollweg. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

57. Machine Learning Focused Data Mining: Finding Anomalies in Code
Kathryn N Rodhouse, Sandia National Laboratories

Auditing source code for bugs and flaws is still largely a manual process (i.e., eyes on code). Manual review becomes intractable as software continues to grow and be updated through increasingly rapid version release processes. We address this by using a machine learning (ML) and data mining methodology to provide a contextually focused flaw identification capability, allowing auditors to make more efficient use of their scant time and attention.

We begin with the frequent itemset mining methodology presented by Li and Zhou in 'PR-Miner: automatically extracting implicit programming rules and detecting violations in large software code', developed to find rule violations in large source code bases. Its core hypothesis is that programs in large, well-developed code-bases follow implicit rules, and that violations of those rules can lead to flaws (e.g., when a call to 'lock' is not followed by a call to 'unlock'). Frequent itemset mining is used to identify such implicit rules and their violations, and specifically in the case of Li and Zhou, is applied to triaging high-confidence, high-support anomalies for auditors. We will show that while Li and Zhou were able to identify strongly anomalous flaws, their methodology is unable to capture medium-confidence, low-support rules (i.e., nuanced rules that are commonly followed but do not occur as frequently as calls to 'lock' and 'unlock'). Finding violations of such rules requires a different, more focused approach.

The first step in our work was to develop the ability to focus the frequent itemset mining technique only on parts of the code base that have a specific functionality, on the theory that each sort of functionality has a distinct set of implicit rules. As an example, we built ML models to detect memory handling functions to mine rules from, as incorrect memory handling can lead to program flaws: corruption of valid memory (overwriting pointers), inappropriate memory accesses (double free), and memory leakage (never free).

We next addressed appropriately scoping the source code boundaries, as this is a critical step in creating itemsets. Li and Zhou proposed creating an itemset for each function within a source code base; however, functions may serve multiple purposes (e.g., manipulating multiple types of data), conflating what is contained within a function-level itemset. To scope the itemsets, we explored program graph based itemsets that reduced the focus of an itemset from a function to a sub-function defined by program graph relationships.

Using the predictive output of the memory handling models and the program graph based itemsets, we show that ML focused frequent itemset mining provides an enhanced capability in detecting violations of medium-confidence, low-support rules.

Joint work with W. Philip Kegelmeyer, Sandia National Laboratories. SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.

58. Embracing Randomness for Uncertainty Quantification in Neural Networks
Ahmad A. Rushdi, Sandia National Laboratories

Neural network models have attracted a lot of research attention in Scientific Machine Learning (SciML) problems. However, they tend to be overconfident when reporting typical point-estimate predictions in classification and regression problems. This could be very harmful when dealing with costly numerical simulations or high-stakes decisions in national security applications. In this work, we explore uncertainty quantification techniques in order to assess the quality of neural network predictions. To understand their variability, we rely on different sources of randomness associated with training samples, data observation order, weight initialization, dropout methods, and ensemble formations. Motivated by typical SciML situations, we assume a limited sample budget and suggest approaches for reporting and possibly reducing uncertainty.

59. Sequential Design for the Calibration of Materials Strength Models
Kathleen Schmidt, Lawrence Livermore National Laboratory

Due to the time and expense associated with physical experiments, the next experimental conditions should be chosen to optimally inform parameter estimation. In the case of material strength, the type of experiment performed -- and, thus, the type of data collected -- depends on the choice of strain rate. Here, we consider two different types of data: stress-strain curves from low-strain rate experiments and deformed Taylor cylinder profiles from higher-strain rate experiments. We employ mutual information, based on Shannon entropy, to select which experiment should be performed to obtain the greatest reduction in strength model parameter uncertainty for tantalum.

LLNL-ABS-801260. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

60. Development of automated pipeline for time-resolved link-wise vehicular energy consumption in the Chattanooga, TN road network

Joseph Severino, National Renewable Energy Laboratory

The Department of Energy (DOE) has shown strong interest in detecting energy inefficiencies in regional road networks, so as to derive energy consumed at a high spatial temporal resolution. We have developed a workflow to automate the estimation of time-resolved vehicular energy consumption over each link in a road network of interest. The road network used in the current work is centered around the city of Chattanooga, Tennessee and its bordering regions. Utilizing the most mature road network for the Chattanooga, TN region, vehicle speed & count data from TomTom in conjunction with machine learning methods, we have developed an automated pipeline to estimate energy consumption for every link in the network. The first step in the pipeline is ingesting vehicle probe counts and speed estimates from TomTom API. In the next step, the probe counts, speed profiles and other exogenous data (i.e. road types, weather data, ground-truth volume counts and more) were used as input to a supervised learning algorithm [1] to estimate the number of vehicles throughout the entire region for each road segment. These volume estimates were then mapped to a unified road network that contained additional important information such as percentage change in gradient across a link, number of lanes and link lengths that are features in pre-trained single vehicle energy-consumption models available with the RouteE software developed at NREL [2]. The per vehicle energy consumption on each road link predicted using appropriate RouteE vehicular models were multiplied by the volume estimate for the corresponding link over a given time period to predict energy consumed per link for the time interval of interest. Currently, work is underway to improve both the RouteE per vehicle energy estimate and the volume estimates derived from TomTom probe counts. We have also explored the correlation of the link-wise energy estimates with the features of the pre-trained RouteE machine learning model in order to gain insight into what factors contribute most to the link-wise energy consumption.

Joint work with Ambarish Nag (NREL), Yi Hou (NREL), Juliette Ugirumurera (NREL), Wesley Jones (NREL) National Renewable Energy Laboratory (NREL).

[1] Hou, Yi, Young, Stanley E, Dimri, Anuj, and Cohn, Nicholas. Network Scale Ubiquitous Volume Estimation Using Tree-Based Ensemble Learning Methods. United States: N. p., 2018. Web.

[2] Holden, Jacob, Harrison Van Til, Eric Wood, Lei Zhu, Jeffrey Gonder, and Matthew Shirk. 'Trip Energy Estimation Methodology and Model Based on Real-World Driving Data for Green-Routing Applications.' Transportation Research Record 2672, no. 24 (2018): 41-48.

61. Data Analytics Applied to Coal Fired Boilers for Detecting Leaks

Larry Shadle, National Energy Technology Laboratory

The increased reliance upon variable energy resources for power generation requires coal-fired power plants to operate in load following or cycling modes. These dynamic operations increase the occurrence of thermal stresses in boilers that produce steam leaks; boiler tube leaks are the primary cause of forced outages in the existing coal fired units. Process data from two coal-fired power plants were analyzed using to evaluate the potential to use data analytics methods to detect tube leaks. The power plants both experienced forced outages to repair leaks in the water wall region of these heat exchangers. The operational effects of these leaks in an 800 MWe super critical boiler were compared to similar located leaks in a 300 MWe sub critical boiler system. Multivariable statistical analyses, including Principal Component (PCA), Cluster Analysis (CA), and Fischer Discriminant Analysis (FDA) were used to characterize process operations with and without the steam leak. Discriminant functions were compared and analyzed for similarities and differences between the plants. Three operational states of the boiler were studied including: 1) normal operations a year prior to the leak (Set A), 2) normal operations immediately before the leak (set B), and 3) the operations with the leaks (Set C). The original datasets were then classified into three different groups; one of the groups was assigned as training while the rest were reserved for validation. PCA was used to reduce dimensionality of the original datasets. The CA and FDA were used to investigate the relationship between process variables with the loads and to discriminate among faults and categorize the processed data into two main groups: non-leak (0) and leak (1). More than 12,000 observations were analyzed, and misclassification were quite few. The analyses also revealed that the different normal operations could also be accurately separated into different groups.

Joint work with Natarianto Indrawan, LSRT/NETL; Rupendranath Panday, LSRT/NETL; and Umesh Chitnis, ABB.

62. Using Wearable Sensor Data to Infer Social Groups and Influence in an Organization
Benjamin Sims, Los Alamos National Laboratory

Wearable sociometric sensors are an increasingly important tool for understanding organizational dynamics, enabling close tracking of behavior in time and space that is not possible through other methods. We analyzed data shared with us by a research team led by the University of Notre Dame as part of the IARPA MOSAIC initiative. For this analysis, we looked at a cohort of 26 participants from a single workplace, followed longitudinally during the year 2018. Participants were instrumented with fixed (home, office) and mobile (keychain, wallet) beacons, which emitted Bluetooth RF signals that enabled real-time tracking of proximity of participants. Participants also completed baseline and daily surveys to track affect and behavior, which we refer to as 'ground truth.' We analyzed beacon data, using close proximity as a proxy for interaction, to assess dynamic composition of collaborative groups over time and to quantify influence among co-workers.

Joint work with Riccardo Boero, Peter Hraber, Kim Kaufeld, Lissa Moore, Ethan Romero-Severson.

63. 3D-Resnet for identification of sample orientation
Jan Strube, Pacific Northwest National Laboratory

Forecasting of material failure and prediction of material properties is an active area of research in additive manufacturing. We present our work on using machine learning techniques to classify 3D-printed samples. We analyzed computer tomography images of various samples to categorize the orientation in which the samples were printed. We demonstrate that our implementation of a 3D convolutional neural network based on the classical Resnet architecture can effectively identify the sample orientation. To help validate the network performance we have developed activation maps that highlight the regions of the data that have a large significance for the network output. We present how we use this activation map in the validation of the network.

64. Reinforcement Learning to Guide Block Copolymer Self-Assembly
Christine Sweeney, Los Alamos National Laboratory

This poster details a reinforcement learning solution to the problem of guided self-assembly of block copolymers. Block copolymers have potential uses in nanotechnology applications, photolithography, photonics, and controlled drug delivery. We describe how this problem is mapped to the interactive learning paradigm of reinforcement learning and details the implementation in both 2D and 3D. Results from small-scale reinforcement learning and preliminary results on large-scale reinforcement learning are presented. The aim is to scale this problem to a higher level of complexity and ultimately make the learned policy available to experimental scientists in hopes of drastically reducing the experimental and computational resources expended to discover an efficient path to creating block copolymer structures of interest.

Block polymers may be combined to create functional biomaterials useful for catalysis, sensors, nanotechnology and renewable energy. In experiments at Department of Energy light source user facilities, scientists use a laser beam to guide the self-assembly of block copolymers from disorder toward target morphologies such as horizontal tubes. However, the assembly process is nonlinear, resulting in a pathway through a solution space that is complicated and requires control over the steps instead of a simple linear path. This process can take hundreds of experimental trials to achieve the target morphology.

Reinforcement learning is a machine learning technique used for control problems where interaction is required in order to learn a policy to reach a target state. The problem of guiding temperature during self-assembly of block copolymers can be formulated as a finite Markov Decision Problem, which lends itself to a reinforcement learning solution. We implement fast 2D and 3D partial differential equation-based simulations for the 'environment' to inform a learning 'agent' about the physics of this process in reinforcement learning. The agent can either 'exploit' the policy learned so far or 'explore' new avenues available in the environment's parameter space. The agent's goal is to maximize their long-term reward of achieving a target morphology.

Our reinforcement learning system is implemented for both 2D and 3D block copolymer simulations and is able to do 'online' learning, which enables agents to make explore versus exploit choices during training episodes. This has been implemented on a small number of nodes on the pre-exascale platform Summit and preliminary results are also available for larger numbers of nodes.

Joint work with Frank Alexander, BNL Anthony DeGennaro, BNL Sayan Ghosh, PNNL Yunzhi (Lucy) Huang, PNNL Ai Kagawa, BNL Zexi Lu, PNNL Jamal Mohd-Yusof, LANL Vinay Ramakrishnaiah, LANL Malachi Schram, PNNL Brian Van Essen, LLNL Paul Welch, LANL Kevin Yager, BNL Shinjae Yoo, BNL. LA-UR-20-20630

65. Distributed heterogeneous compute infrastructure for the study of additive manufacturing systems
Mathew Thomas, Pacific Northwest National Laboratory

We present a computing and data science effort at PNNL for an end-to-end analysis pipeline for chemical dynamics studies using High-Performance Computing (HPC) resources. The current computing model uses DIRAC (Distributed Infrastructure with Remote Agent Control) for its workflow and data management capabilities. The DIRAC project is a complete Grid solution that provides all the necessary components to create and maintain distributed computing systems. It forms a layer on top of other computing infrastructures, which isolates users from the direct access to the computing resources and provides them with an abstract interface hiding the complexity of dealing with multiple heterogeneous services. A detailed meta-data assignment using the DIRAC File-Catalog is used to automate the stages of data processing. The DIRAC system is deployed on containers managed using a Kubernetes cluster to provide a scalable infrastructure. A modified DIRAC agent provides the ability to submit jobs using singularity on dedicated and opportunistic HPC sites. The data products from this pipeline are feed into a graphics processing unit (GPU) cluster that runs various Machine Learning (ML) tasks, such as 3D convolution networks and/or physics aware temporal models.

Joint work with Malachi Schram, Kevin Fox, Noah S. Oblath, Elvis Offor, Zach Kennedy, Chris Barrett, Anil K. Battu, Tamas Varga.

66. Visual Analytics Science and Technology (VAST) Challenges - Emergency Response
Sandy Thompson, Pacific Northwest National Laboratory

The IEEE Visual Analytics Science and Technology (VAST) Challenge provides an innovative venue to engage visualization, visual analytics, and data scientists across academia, government and industry to apply their latest visual analytics R&D to a problem of direct interest. PNNL has developed, fielded, and executed the IEEE VAST Challenge for twelve years, during which time it has garnered international acclaim, thousands of dataset downloads, and submissions by organizations in 23 countries. The 2019 challenge focused on nuclear emergency response, helping a city grapple with the aftermath of an earthquake that damaged their nuclear power plant. Mini-challenge 1 asked researchers to integrate seismic and open source information to understand damage and prioritize emergency response. Mini-challenge 2 looked at crowd sourced sensors together with stationary city sensors to understand the spread of radioactive contamination from the damaged power plant. Mini-challenge 3 looked to turn social media posts into an information source helping the city where to focus efforts and understand the concerns of the populace. Together these three challenges were then brought together to ask the analytics community to answer the broader question of what should the city do.

Joint work with Jereme Haack, Kris Cook.

67. Classification of light and heavy-duty vehicles from live traffic cameras
Austin C. Todd, National Renewable Energy Laboratory

The transportation sector is the largest contributor to greenhouse gas emissions in the United States, accounting for about 29% of total greenhouse gas emissions. The dominant source of transportation-related emissions (over 80%) is from light- and heavy-duty vehicle traffic. Therefore, reducing congestion and improving traffic mobility can greatly reduce the country's total greenhouse gas emissions. Researchers from Oak Ridge National Laboratory and the National Renewable Energy Laboratory are assessing ways to modify traffic control strategies to improve the regional mobility through a test city of Chattanooga, TN. A key requirement for this work is to obtain accurate counts of total light and heavy-duty vehicles such that we may compute reasonable energy consumption estimates and identify successful control strategies based on vehicle type. Therefore, a method to determine the light and heavy-duty vehicle counts would greatly improve our impact on the City of Chattanooga. Since complete estimates of light and heavy-duty vehicles are rarely available, we have implemented a method to directly detect vehicles from low-resolution roadside traffic video cameras using the You-Only-Look-Once (YOLO V3) deep image processing network. We highlight many of the computational challenges associated with performing object detection in low-resolution, operational, live video streams. These challenges include image degradation during low visibility events (e.g. rain, nighttime) and camera movement by traffic control operators. Our preliminary results are compared to total aggregated volume estimates from radar detection sensors adjacent to camera locations, and we demonstrate the utility of using an object detection algorithm for performing vehicle counts. Rough estimates of the breakdown between light and heavy-duty vehicle indicate about 10% of highway traffic along our test corridor are heavy-duty vehicles. Heavy-duty vehicle traffic reaches peak percentage of 16% of all traffic during the 10 o'clock hour and a peak total volume during the early PM peak (3-4pm). This presentation will also outline our continuing plans to expand our implementation from one test corridor to the entire Hamilton County transportation region.

68. Data Analytics in Smart Manufacturing

David M. Tralli, The Aerospace Corporation

The Aerospace Corporation is a member of the Clean Energy Smart Manufacturing Innovation Institute (CESMII) awarded to the Smart Manufacturing Leadership Coalition under a DOE Cooperative Agreement in 2016. CESMII is a national network that brings together over $140 million in public-private investment and more than 200 partners from leading manufacturers and universities in over 30 states. CESMII solicits projects through RFPs to accelerate Smart Manufacturing (SM) adoption through integration of advanced sensors, data analytics, platforms and controls to improve precision, performance, quality, sustainability and energy management. The CESMII SM Platform provides the enabling infrastructure and Marketplace for Information Technology and Operational Technology integration (i.e., cyber-physical systems or CPS) to optimize manufacturing process and production efficiency. The SM Platform is an industry-specified and managed assembly of web-accessible infrastructure and cloud services that function as a shared operating system for multiple application system composability, infrastructure service, and deployment integration and interoperability with Marketplace access. The CESMII Project Portfolio tackles R&D challenges and knowledge gaps in CPS: security, multi-sensor data fusion, sensor-actuator-human interfaces, process model V&V and uncertainty; data structures, contextualization, and configuration management; and reference architectures and platforms for process technology digitization. In Data Modeling for Machine Learning and Data-Centric Analytics, CESMII projects address secure, scalable, and interoperable on-premise, edge, and off-premise cloud technology with reusable data analytics, automated workflows and toolkits for ML applications from supervised to deep learning. With the high capital cost of transitioning from largely manual operations to an automated environment, the goal of the Smart Connected Workers in Advanced Manufacturing Project, led by UC Irvine, with partners Atollogy, Aerospace, General Mills, Honeywell, San Diego Supercomputer Center, and CSU Northridge, is to develop design tools and open source reference architectures that enable reliable assessment and implementation of data flows for affordable, scalable, accessible and portable (A.S.A.P.) SM systems. The Smart Connected Worker System utilizes vision-based technology to characterize workflows to be effectively combined with sensor data. Simple closed-loop program models monitor energy consumption, processing parameters, and work efficiency. An ML routine can learn when it is most likely that an energy fluctuation caused a significant defect in a laser-processed part, for example. Aerospace contributions are highlighted, associated with its Design & Manufacturing Hub, guiding the strategy of internal R&D in composites, microelectronics, metals and fabrication for the Defense & Aerospace industry.

69. The Multifaceted Relationship Between Simulation and Data in Metal Additive Manufacturing Process Modeling

John A. Turner, Oak Ridge National Laboratory

The Exascale Computing Project (ECP, https://exascaleproject.org/) is a U.S. Dept. of Energy (DOE) effort developing hardware, software infrastructure, and applications for computational platforms capable of performing 10^18 floating point operations per second (one ‘exaop’). The Exascale Additive Manufacturing Project (ExaAM) is one of the applications selected for development of models that would not be possible on even the largest of today’s computational systems. In addition to ORNL, partners include Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), the National Institute for Standards and Technology (NIST), as well as key universities such as Penn. State Univ., Purdue Univ., UCLA, Univ. of Penn., Univ. of S. Carolina, and Georgia Tech.

Although the primary focus of the project is on a traditional 'forward simulation' workflow (highly-resolved melt pool simulations to inform part-scale finite element thermomechanics simulations, drive microstructure evolution, and determine constitutive mechanical property relationships based on those microstructures using polycrystal plasticity), multiple collaborations are under way to explore opportunities to both benefit from and contribute to more data-centric approaches.

For example, a connection to the Aeolus project (https://aeolus.oden.utexas.edu/), one of the DOE Office of Science Advanced Scientific Computing Research (ASCR) Mathematical Multifaceted Integrated Capabilities Centers (MMICC) is integrating machine learning techniques via the 'Lift and Learn' approach to generate reduced order models for lower length scale phenomena. Related efforts are under way via collaborations with researchers at Georgia Tech. and within the ECP ExaLearn co-design center.

Ultimately, we view ExaAM as a project that can both benefit from data-driven approaches (such as accurate reduced-order models for lower length scale phenomena) and contribute to efficient AM process models necessary for applications such as real-time control and topology optimization.

Joint work with Matthew Bement (ORNL), James Belak (LLNL), Chris Newman (LANL). The core ExaAM team includes James Belak, co-PI (LLNL), Nathan Barton (LLNL), Matt Bement (ORNL), Neil Carlson (LANL), Robert Carson (LLNL), Jean-Luc Fattebert (ORNL), Neil Hodge (LLNL), Zach Jibben (LANL), Brandon Lane (NIST), Lyle Levine (NIST), Chris Newman (LANL), Balasubramaniam Radhakrishnan (ORNL), Matt Rolchigo (LLNL), Stuart Slattery (ORNL). This work was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.

70. Integrating Physics Knowledge in Multi-Sensor Machine Learning Models
Charlie Vollmer, Sandia National Laboratories

Can we improve predictions derived from noisy, sparse, and high-dimensional data by applying physically informed constraints to machine learning models? We hypothesize that by incorporating scientifically informed constraints within purely data-driven machine learning models, we can improve the prediction and scientific inference for sparse data that are too expensive to collect and/or the design space is prohibitively large for the collection of a sufficient training set. By incorporating domain knowledge, the developed model will include scientific information and require less data to achieve state-of-the-art results. We focus on applying our methods to a material classification problem using a training set of images generated from hyperspectral computed tomography (H-CT). Each scan produces gigabytes worth of data which, when considering the combinations of potential effect variables, can make obtaining a sufficient number of replicates for a purely data-driven approach infeasible. We instead leverage a priori knowledge of the expected spectrum to inform our predictions.

71. Machines Learning about Machines - ML for Analysis & Control of HPC Infrastructure
Felix Wang, Sandia National Laboratories

As high-performance computing (HPC) systems have scaled to enable the complex computational demands of scientific computing, not surprisingly HPC systems have grown increasingly complex. Beyond simply the number of integrated components, this complexity arises due to heterogenous approaches being employed to enable a variety of workloads as well as increasingly sophisticated algorithms. Altogether, these advances impose considerable challenges upon system software striving to maximize system utilization. This includes scheduling challenges such as how jobs are allocated across nodes and the order to execute jobs, as well as operational resources such as the clock frequencies nodes operate upon. Not only can optimizing these resource allocation tasks increase the computational workloads HPC systems may perform, but it can also yield significant energy or run-time savings. Here we highlight utilizing machine learning both to enhance the understanding of system performance as well as identifying how machine learning can help to automate resource management. Our data analysis examines historical job runs on the Sandia National Labs Astra ARM based supercomputer focusing upon trends in power consumption as a function of job execution. We use this analysis to identify techniques by which machine learning can further automate resource allocation with an example application of adjusting clock frequencies to attain power savings while minimizing run time impact.

Joint work with Sam Green, Kevin Pedretti, Craig M. Vineyard, Andrew Younge.

72. Modeling multivariate censored data: an application to assess base levels of Beryllium
Richard Warr, Brigham Young University

The Department of Energy (DOE) is required by law to ensure that its employees work in an environment safe from hazards such as beryllium (Be), a metal that, if inhaled, can cause scarring and inflammation in the lungs. This requirement involves regularly collecting data to demonstrate that the amount of hazards present in the work environment is less than a specified OSHA limit, e.g., 0.1 micrograms per square cm for Be. Data are collected at a suspected site using pre-moistened ‘wipes,’ which are sent to an independent laboratory where content is measured for Be and other toxins. In a clean environment (i.e., only naturally occurring Be), Be is typically below both the OSHA limit and the measurement system’s detection limit. If Be is below the detection limit, the limit is returned, producing a left-censored measurement. Other metals are also collected from the swipe sample, such as yttrium, aluminum, and vanadium which are also subject to censoring. This work illustrates how to estimate a joint distribution, when the variables are subjected to censoring.

Joint work with Brian Weaver (LANL) and Kimberly Kaufeld (LANL).

73. Identification of Potential Crack Oriented Neighborhoods from Template Feature Matches
Jim Wendelberger, Los Alamos National Laboratory

Confocal microscope images of the inner surface of a hazardous material container contain features of potential interest to determine container integrity. These features are matched to various templates that represent various crack attributes. The matching templates provide specific locations of likely potential crack features. These specific locations are used to identify the most likely neighborhood or area of locations on the image that may contain a potential crack feature. The most likely area is an oriented neighborhood. The oriented neighborhood may be displayed along with the original image to aid evaluation of the region for potential cracks of interest by a subject matter expert.

74. A Data Driven Operational Model for Traffic and Freight Analysis at Dallas Fort-Worth Airport
Lindy Williams, National Renewable Energy Laboratory

Airports are on the front lines of significant innovations allowing the movement of more people and goods faster, cheaper, and with greater convenience. As air travel continues to grow, airports will face challenges in responding to increasing passenger vehicle traffic, which leads to lower operational efficiency, poor air quality, and security concerns. This poster introduces and evaluates data-driven methods for traffic demand forecasting, which will allow airport operations staff to accurately forecast traffic and congestion using machine learning-enhanced modeling. Using detailed data describing individual vehicle arrivals and departures, aircraft movements and weather at Dallas Fort-Worth (DFW) International Airport, we evaluate multiple real-time forecasting methods including seasonal ARIMA, Prophet, and modern supervised machine learning algorithms. We also apply a similar approach for freight movement on the ground, comparing and contrasting the similarities and differences between these two different airport systems. This framework and analysis can assist airport and seaport transportation hubs like DFW, as they tackle daily operational challenges and decisions for complex systems, as well as explore the integration of emerging technology and expansion of their services into long term plans. This work also demonstrates how scientific machine learning can be applied to operational challenges through the development of integrated decision support and intelligent automation.

Joint work with Monte Lunacek (NREL), Joseph Severino (NREL), Karen Ficenec (NREL), Juliette Ugirumurera (NREL), Matthew Eash (NREL), Yanbo Ge (NREL), Melissa Queen (UW), and Caleb Phillips (NREL).

75. Granulemeta: file-level metadata extraction for Earth science data applications.
Bruce E. Wilson, Oak Ridge National Laboratory

The pace of data acquisition is accelerating across science domains, forcing data distributors to adapt their approaches to managing ballooning volumes and varieties of data. One essential function of data distribution systems is the collection of high-quality metadata that describe data files. The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC), a distributor of Earth science data collected for missions sponsored by NASA's terrestrial ecology program, has developed a tool for extracting the comprehensive metadata required to support the high-level data applications available through NASA's Earth Science Data and Information System (ESDIS). Data stewarded by the ORNL DAAC are unique in their varieties, including observations from ground, airborne, and satellite platforms; outputs from diverse models; and data from lab notebooks and data loggers. Granulemeta automatically extracts metadata such as spatial and temporal extent, variable names and units, and basic statistics from a batch of data granules -- an ESDIS term equivalent to files -- and organizes them into structured outputs that meet the requirements of the ORNL DAAC's data management system. The tool accesses metadata through libraries specific to each data format and relies only on Python 3 with a minimal set of dependencies. Granulemeta is capable of parsing arbitrarily-formatted tabular data stored in plain text formats to record characteristics such as header size, column names, and delimiters, and uses regular expressions to identify columns that store information about the time and location of observations. It leverages the Python interface to the Geospatial Data Abstraction Library (GDAL), and its sister project the OGR Simple Features Library (OGR), to access metadata stored in common geospatial data formats, including HDF, netCDF, GeoTIFF, ESRI Shapefile, KML/KMZ, GeoJSON, GeoPackage, and many more. Its capabilities are best suited to data sets stored in the self-describing netCDF format following the Climate and Forecast Conventions, which indicate exactly where important metadata are stored within a file. Extracted metadata can be mapped to custom outputs with little effort. Granulemeta is currently being tweaked for use by other projects with similar metadata requirements. Granulemeta can access metadata from files in remote data stores, including Amazon S3, and is extendable to meet the ORNL DAAC's ever-changing needs.

Joint work with Jack McNelis, Yaxing Wei, Tammy Walker, Debjani Singh, Daine Wright, Chris Lindsley, Vinny Palandro (all of Oak Ridge National Laboratory).

76. Scalable spatio-temporal modeling using a fast multipole method for 3D tracer concentration breakthrough data with magnetic resonance imaging

Hongkyu Yoon, Sandia National Laboratories

With recent advances in sensing technology, large volumes of hydrogeophysical and geochemical data can be obtained to achieve continuous tracking of the movement of a fluid or a plume in the subsurface. However, characterization with such a large amount of information requires prohibitive computational and storage costs associated with matrix construction, matrix-matrix multiplication, and linear system solution. To tackle such challenges, we present a spatio-temporal modeling without explicit construction of the covariance matrix, and take advantage of the parallel black-box fast multipole method (FMM) and the parallel inverse Fast Multipole Method (IFMM) for matrix-vector multiplication and linear system solution, respectively. Overall, our approach requires O(N) computation and storage. For an illustrative example, we use 6 million transient tracer concentration measurements in a laboratory-scale 3-D sand-box obtained using magnetic resonance imaging to monitor real-time tracer plume migration. The sand-box was filled with discrete patterns of 5 different sizes of sands at 1 cubic centimeter, creating relatively heterogeneous patterns of permeability distribution and associated complex transport of the tracer. We demonstrate that the spatio-temporal modeling can be performed with a big data set such as this concentration data for real-time plume tracking.

Joint work with Jonghyun Lee, University of Hawaii at Manoa, Chao Chen, University of Texas Austin, Toru Takahashi, Nagoya University, Eric Darve, Stanford University. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.