Exploring Mars with the Curiosity Rover: Worth Every Penny at a Million Dollars a Day
Nina L. Lanza, Los Alamos National Laboratory

Since her arrival at Mars on August 5, 2012, the Curiosity rover has been exploring the martian surface and providing exciting new insights about the geological history of that planet. The goal of this mission is to determine the past and present habitability of Mars. As such, the search for water, hydrated minerals, and organic molecules is particularly important. Over the past three years, Curiosity has discovered abundant evidence that her landing site, Gale crater, once contained persistent liquid water in the form of a lake and numerous stream systems. These scientific investigations are made possible by the 10 rover payload instruments, which provide a wide range of data products including spectra and images. One of these instruments is ChemCam, which was designed, built, and led by Los Alamos. ChemCam consists of two instruments, a laser-induced breakdown spectroscopy (LIBS) instrument and a Remote Micro-Imager (RMI) camera. The LIBS instrument obtains chemical information about targets up to 7 m from the rover, while the RMI provides a geologic context image for the returned chemistry data. To date ChemCam has obtained compositional information for ~300,000 locations on over 4,000 geological targets on Mars, an unprecedented amount of data for a single payload instrument. The knowledge gained from ChemCam on the Curiosity mission will be applied to the upcoming Mars 2020 rover mission, for which Los Alamos has been selected to build another instrument called SuperCam.

Mathematical Tools for High Resolution 3D X-ray Movies
Dula Parkinson, CAMERA, Lawrence Berkeley National Laboratory

3D scanning a sample is cheap—there's an app for that! But a 3D movie of the interior of a moving object with micrometer definition is not so cheap or easy. To do it well, you need an intense X-ray source to see inside the object with good resolution (~$1 billion [1],[2]), a high performance computing system to handle the data (~$70 million[3]), and advanced analysis tools (~$10 million[4]). We will show some of these not-so-easy 3D movies and the accompanying analysis done at Lawrence Berkeley National Laboratory (LBNL), with data collected at the microCT instrument of the Advanced Light Source (ALS), analyzed at the National Energy Research Scientific Computing Center (NERSC) with tools from the Center for Advanced Mathematics for Energy Research Applications (CAMERA).

The 3D movies to be featured show geological, biological, and composite materials changing due to their environment (heat, compression, moisture, etc.). Understanding the changes requires not just visualizing the data, but identifying and quantifying features of interest and tracking them in time. We take a multi-pronged approach to the challenge of automating this analysis for the huge amounts of data. Approaches include separating the phases that compose the materials using a parallel Markov Random Field method (PMRF) and using a GPU accelerated 3D fast filtering package (F3D) and pattern matching to detect and measure specific features in the samples. Results show the accuracy of the methods and the increase in speed and ease-of-use over previous methods.

Join work with Talita Perciano, Dani Ushizima, Hari Krishnan, Jamie Sethian

[1] https://www.bnl.gov/ps/nsls2/about-NSLS-II.php
[2] https://www-als.lbl.gov/index.php/about-the-als/quick-facts.html
[3]http://cs.lbl.gov/news-media/news/2014/nersc-cray-intel-to-collaborate-on-next-generation-supercomputer/
[4]http://newscenter.lbl.gov/2015/09/22/new-support-for-camera-to-develop-computational-mathematics-for-experimental-facilities-research

---

Dark Energy Survey
Douglas Tucker, Fermilab

The Dark Energy Survey (DES) is an astronomical optical imaging survey of one quarter of the Southern sky. The survey is being conducted with a 3 square degree wide-field mosaic camera (the Dark Energy Camera, or DECam) on the Blanco 4-m telescope at the Cerro Tololo Interamerican Observatory in the Chilean Andes, and it has just completed its third year of operations (out of a planned total of 5). The primary scientific goal of the DES is to constrain dark energy cosmological parameters via four complementary methods: galaxy cluster counts, weak gravitational lensing, baryonic acoustic oscillations, and Type Ia supernovae. In this talk, I will discuss DES instrumentation, DES operations, some early DES results, and the benefits of DES and DECam both to the astrophysics community and to the general public.

Inertial Confinement Fusion Implosions at the National Ignition Facility: Experiments and Simulations at Extreme Scales

Brian K. Spears, Lawrence Livermore National Laboratory

The National Ignition Facility (NIF) at Lawrence Livermore National Laboratory strives to achieve fusion ignition using the world's largest laser. This multi-billion-dollar facility produces nearly 2 MJ of laser light to compress a millimeter-radius spherical capsule to a radius of a few tens of microns. Once compressed, the target reaches temperatures exceeding 50 million K and begins thermonuclear fusion of deuterium-tritium (DT) fuel.

The most successful implosions to date have doubled their energy yield due to self-heating by fusion reactions. However, each experimental attempt to improve performance is both expensive (~ $1M) and infrequent (~ weekly). By big data standards, we produce a small number of heavily diagnosed realizations each year.

At the same time, the parameter space for designing a target implosion is vast. Depending how you count, we identify about three dozen parameters that affect the spherical part of the implosion and another few dozen that perturb the shape of the implosion.

To design experiments in this high-dimensional space, we rely on predictive numerical simulations. And, here enters big data. One of our approaches is to explore the topography of parameter space using large ensembles of simulations. Each single simulation is large, taking 4 hours on 48 parallel processors. However, by harnessing the power of the Trinity supercomputer, our workflow system is able to produce about a quarter of a million simulations. When in production, we produce about 1000 simulations per hour, and we require detailed diagnostic post processing at 100 Hz frequencies.

We will give a tour of the typical output from experiments and the comparable output from our simulations. Then, we will turn to our ensemble results demonstrating the global information that we extract and the methods we use for extraction. We'll end with a look forward to ways that we plan to advance our data analysis.

Joint work with Scott Brandon, Michael Buchoff, John Field, Jim Gaffney, Jim Hammer, Michael Kruse, Steve Langer, Ryan Nora, Luc Peterson, Paul Springer

(Some) Directions for Impactful Research at the Intersection of Data Science and the Power Grid

Jean-Paul Watson, Sandia National Laboratories

It is arguable that the term 'big data' is not applicable to power grid control, operations, and planning - particularly in comparison to remote sensing and cyber applications. However, power grid applications should still be of significant interest to the data science community, due to the potential and largely differentiating linkages between diverse spatio-temporal data sources (e.g., forecasted and realized renewables output) and control algorithms used to operate power grids worldwide. In particular, the outputs of data science analyses can both directly impact existing power grid operational paradigms and inform the design of novel control architectures and investment plans, ultimately improving cost-effectiveness, reliability, and resiliency.

In this talk, I will introduce various exemplars of existing, on-going research projects that lie firmly at the data science/power-grid intersection. These projects range from being loosely to tightly coupled with power grid control and planning paradigms, and range from (physics) 'model-free' approaches to more complex applications that directly couple data science outputs with control models and algorithms. Typically, such projects are staffed almost exclusively by engineers and operations researchers - including the speaker, who have leveraged and developed a range of heterogeneous and often ad-hoc approaches to dealing with the requisite data analyses. Direct engagement with the data science community has the potential to dramatically improve the fidelity and rigor of data modeling tasks (e.g., projecting hurricane impacts) and the effectiveness of control architectures (e.g., operations models based on probabilistic forecasts of renewables generation). Further, I argue that major impediments to deploying advanced control technologies in the power grid often center around data science questions - for which we often have poor or inadequate answers.

Graphs and the Power Grid
Emilie Purvine, Pacific Northwest National Laboratory

In this talk I will describe work being done for the DoE-funded Center for Multifaceted Mathematics for Complex Energy Systems (M2ACS). In this center we are developing new mathematical techniques that have applications for complex systems including the power grid. This is a large interdisciplinary project involving both mathematicians and power engineers working together to solve complex power grid inspired problems. My work in particular is focused on how graphs can be used to both model and understand the complexities of a power grid. Random graph models for the power grid are difficult to create since it is such a specifically engineered system. However, due to the proprietary nature of most systems, researchers need a way to test algorithms on graphs that resemble the power grid to some statistical accuracy. The first focus area of our work has been to discover a layered, network-of-networks decomposition of known power grids from around the world which can be modeled using a random process. Then, our second focus area is in model reduction for power grid simulations. Power systems are often very large and complex but when many elements (e.g., buses and generators) behave similarly or are located close to each other, we may be able to simulate an approximation to the whole system through a reduced order model. My work has been on using graph-based methods like spectral clustering and Fiedler partitioning to group power generators based on behavior and then comparing the simulated system to the full system. I will report on our findings and compare to other model reduction methods.

Data-Driven VERDE: Visualizing Energy Resources Dynamically on Earth
Mallikarjun (Arjun) Shankar, Ph.D., Oak Ridge National Laboratory

We discuss our approach to monitor and present the status of energy infrastructure resources in a user-friendly manner. By combining multiple information sources using direct acquired status data as well as information assembled across different modalities, we create a holistic view of the energy grid infrastructure. Such a broad-based and wide-area view is invaluable in assisting operators and analysts to develop response strategies because it allows them as end-users to combine precomputed and ingested data with their own cognitive context. We show how the VERDE framework offers a flexible pipeline not only to acquire, align, and present data, but also to provide automated predictive analytics services. These analytic services include the use of graph-based interdependency exploration and pattern extraction.

The Wind Integration National Dataset (WIND) and Solar Integration National Dataset (SIND) Toolkits
Bri-Mathias Hodge, National Renewable Energy Laboratory

Renewable power resources (such as wind and solar) are both variable and uncertain in their power output. This presents a number of issues in their efficient integration into the power system. Often simulation studies (known as integration studies) are conducted to gauge the impacts of high penetrations of renewable energy on the electricity grid. These studies often require high fidelity data on wind and solar power output that is not yet available from installed systems. The WIND and SIND Toolkits provide this temporally and spatially granular simulated power output from wind and solar power at over 100,000 sites throughout the continental United States.

Joint work with Caroline Draxl, NREL, Andrew Weekley, NREL, Andrew Clifton, NREL, James McCaa, Vaisala

C. Draxl, A. Clifton, B. Hodge, J. McCaa. The Wind Integration National Dataset (WIND) Toolkit, Appl Energy, 151 (2015), pp. 355-366.

A Multiscale Integrated Modeling Framework for Climate Science and the Multisource Data Challenges
L. Ruby Leung, Pacific Northwest National Laboratory

Climate change is a complex problem that intersects multiple disciplines and processes across a wide range of spatial and temporal scales. Traditionally different modeling systems and approaches have been used in climate prediction, climate mitigation, and impacts, adaptation and vulnerability research. However to inform decisions for the next decades, new modeling frameworks must recognize the co-evolution of the Earth and human systems or interactions between climate, mitigation, and adaptation. The Platform for Regional Integrated Modeling and Analysis (PRIMA) is a multiscale integrated modeling framework consisting of Earth System Models, Integrated Assessment Models, and Sectoral Models that can be flexibly configured to answer specific climate science questions. This presentation will provide an overview of PRIMA, highlight examples of numerical experiments using PRIMA to advance understanding and prediction, and discuss future challenges.

A Frequentist Approach to Multi-Source Classification
Katherine Simonson, Sandia National Laboratories

The classification of unknown entities based on measured data is a fundamental challenge across applications as diverse as medical diagnostics, treaty monitoring, and electronic fraud detection. In many cases, the data available to support classification decisions arise from multiple sources, each with its own unique signal and noise characteristics. The method to be discussed here, known as Probabilistic Fusion (PF), provides a means to combine multi-source classification information in a manner that is statistically rigorous and accounts for the uncertainties associated with the constituent sources. PF provides final class consistency scores that are readily interpretable within in a Frequentist framework, and allows complete traceability back to the contributing sources. It is particularly appropriate in applications related to high consequence decision support, where training data may be limited, and “black box” classifiers struggle to gain trust and cultural acceptance. The method will be illustrated with examples from several areas of application.

Joint work with Ross Hansen, Tom La Bruyere, Mark Van Benthem, and R. Derek West (all Sandia National Laboratories)

Methods of Imputation for Multi-Source Data
Emily Casleton, Los Alamos National Laboratory

Multi-source data provide a unique opportunity to learn about a single phenomenon from different perspectives. Specifically, various sensors can indirectly monitor a facility for a rare event, and that event will manifest itself in the data streams differently. Informative features often must be extracted from the raw data to provide a uniform time scale, as each stream will often be sampled at its own unique frequency, and to limit the effect of noise of the raw data. In this talk, we will examine several methods of imputing feature values for multi-source data when the original data stream is not observed. Because each sensor is monitoring the same facility, if the extracted features streams are informative about the event of interest, they will be correlated. Thus, the observed streams will provide evidence of the values from a feature stream that is not completely observed. Two imputation methods that leverage this correlation, through ridge regression and a dynamic linear model, will be examined and compared with the imputation of median values. The effect of estimating missing values versus using only fully observed time steps, will also be examined on the analysis method of interest; here the effect on the elastic net classifier.

Big Data, Geospatial Computing, and My 2 Cents in an Open Data Economy
Budhendra Bhaduri, Oak Ridge National Laboratory

In this rapidly urbanizing world, unprecedented rate of population growth is not only mirrored by increasing demand for energy, food, water, and other natural resources, but has detrimental impacts on environmental and human security. Much of our scientific and technological focus has been to ensure a sustainable future with healthy people living on a healthy planet where energy, environment, and mobility interests are simultaneously optimized. Current geoanalytics are limited in dealing with temporal dynamics that describe observed and/or predicted behaviors of entities i.e. physical and socioeconomic processes. With increasing temporal resolution of geographic data, there is a compelling motivation to couple the powerful modeling and analytical capability of a GIS to perform spatial-temporal analysis and visualization on dynamic data streams. However, the challenge in processing large volumes of high-resolution earth observation and simulation data by traditional GIS has been

compounded by the drive towards real-time applications and decision support. Ability to observe and measure through direct instrumentation of our environment and infrastructures from buildings to planet scale, coupled with explosion of data from citizen sensors brings much promise for capturing the social/behavioral dimension and provides a unique opportunity to manage and increase efficiencies of existing built environments as well as design a more sustainable future. This presentation will explore the intriguing developments in the world of Big Data, geospatial computing, and plausible ways citizens can all become part of the open data economy for advancing science and society.

Improving the Trust in Results of Numerical Simulations and Scientific Data Analytics
Franck Cappello, Argonne National Laboratory and University of Illinois at Urbana Champaign

Since the introduction of fault tolerance in HPC, the research has mainly focused on process crashes, network dysfunctions, and radiation induced bit flips. This orientation comes directly from the generic fault tolerance problems and solutions seen in distributed systems. Thus, the main objective has been to protect the execution - and that mostly in an application-agnostic way.

Many real-life examples, however, demonstrate that application results can also be disrupted by hardware and software malfunctions due to aging and bugs. The risk of cyber attack is also taken seriously. Unfortunately, fault tolerance and resilience techniques currently used in HPC have not been designed for these types of disruptions; and in practice most fail to provide useful solutions.

What matter is not only to protect HPC executions; it is also to protect the correctness of the results such executions produce.

The scientific problem behind this statement is the trustworthiness of HPC application results. And to improve the trustworthiness, we must start from the results of the execution, as opposed to how the execution is implemented with processes and communications. The main objective is to avoid influential data corruptions related to the user-expected accuracy.

In this talk, we will review the notion of trust, the different types of disruptions leading to corruption of results, the ways that users build trust in application results, and the limitations of current techniques (fault tolerance/resilience, validation and verification, uncertainty quantification) in the context of scientific computing. We will present examples of results corruptions, some leading to catastrophic consequences, as well as potential approaches to improve result trustworthiness.

Franck Cappello, Emil Constantinescu, Paul Hovland, Carolyn Phillips, Marc Snir, and Stefan Wild, Improving the trust in results of numerical simulations and scientific data analytics, http://www.mcs.anl.gov/publication/improving-trust-results-numerical-simulations-and-scientific-data-analytics

---

Searching for the Needle in the Needle Stack: Situational Awareness and Cyber Security at Enterprise Scale
Rhiannon Weaver, Neya Systems, LLC

Cybersecurity as a field has embraced Big Data in a one-way relationship. Defenders want to collect everything, but are only recently coming to understand that large-scale data collection still needs the right context and analytical scope in order to be useful. The Cyber mission is fundamentally different from other massive data mining efforts that can tailor algorithms to an individual user and look for small-scale feedback, or that can profit from prediction of large-scale trends. The cybersecurity analyst is an aggregator, tasked with investigating alerts from many different individuals (users, computers, IP addresses), which makes false positive costly. The adversary is intelligent, adaptive, and able to enact change at the lowest levels of data granularity, essentially undermining the most basic assumptions that hold trivially in the physical world. As statisticians, we need to help cybersecurity analysts make the most of this complex data with brittle signals, at the largest scale.

This talk is a “lessons learned” overview, compiled from my experiences over 9 years as a statistician and security analyst at CERT, part of the Software Engineering Institute FFRDC at Carnegie Mellon University. I focus on providing a statistical perspective on the landscape of current large scale cybersecurity analysis, its main challenges and where I believe statistical thinking and techniques can provide the most value. The overview is accompanied by a few illustrative examples of the tools and data that we used in research and day to day operations helping to understand and defend carrier-class DoD and Federal/Civilian networks. For over 15 years, CERT has been collecting high-level netflow summaries — who talked to whom, on what channel for how long —

of every single IP-to-IP communication that crosses the border of these networks. Over the last decade, operational analysis has focused on two main efforts: collecting covariate information on IP addresses — WHOIS and DNS — to augment the basic netflow standard, and integrating the results from small-scale malware analysis into large-scale network queries. Both these efforts focus on using simple techniques and subject-matter expertise to provide the right context in which machine learning or statistical models can provide value.

The Hybrid Toolkit for Cybersecurity Analytics
Warren L. Davis IV, Sandia National Laboratories

The Hybrid Toolkit was developed as an outgrowth of a Sandia Laboratories research program aimed at exploring the detection of phishing and spear phishing emails using innovative data analytics. However, the end effect was a much more flexible framework that facilitates collaboration between domain experts and algorithm researchers, and allows innovative research codes to more successfully cross the “Valley of Death.”

The Hybrid Toolkit allows analytics researchers to integrate algorithms into production environments without the burden of overcoming integration issues, in turn resulting in more extensible codes. In addition, it allows domain experts to take advantage more easily of the vast analytics capabilities available from external sources.

This talk will explore the impetus for the Hybrid Toolkit, its architecture, its value to cybersecurity operations, and functionality extensions currently in development.

Joint work with Christopher Nebergall (Sandia National Laboratories), Eunsil Han (Sandia National Laboratories)

Cyberphysical Security
Earl Lawrence, Los Alamos National Laboratory

Much of our most important infrastructure is a cyberphysical system. The power grid is a prime example: it is a physical network governed by the laws of physics and controlled by the cyber network. An attack on the communication network that controls the power grid could not only cause temporary blackouts, but cause permanent damage to generators and transformers. Stuxnet is a well-known example of a cyberphysical attack in which a computer virus was used to destroy Iranian centrifuges. Water systems, gas pipelines, and military bases are other systems that are vulnerable to attacks through the cyber system that have consequences in the physical world. This talk presents work from a project to detect intrusions into such a system. We use data from the climate control system of LANL office building. We treat the problem as anomaly detection in streams of data collected from sensors on the infrastructure, combined with cyber traffic. This talk will present a method for characterizing typical behavior from such streams as a building block for the anomaly detection. We adapt a hidden Markov model used for time series alignment to estimate recurring patterns in multivariate time series. The HMM architecture provides a method for filtering the data as it streams to produce standardized residuals that can be used to detect intrusions.

Data Science, Public Policy, and Social Good: Challenges, Opportunities, and Lessons Learned
Rayid Ghani, University of Chicago

We're all aware of the data science hype right now but turning this hype into reality and actual impact is difficult. In this talk, I'll discuss lessons learned while working on dozens on data science projects over the past few years with non-profits and governments on education, public health, sustainability, economic development, public safety and other social issues. These lessons span from challenges these organizations face when trying to apply data science, to understanding how to effectively build cross-disciplinary teams and teaching them how to do practical data science, as well as what data science and social science research challenges need to be tackled, and what tools and techniques need to be developed in order to have a social and policy impact with data science.

Dirty Data: A Data Science Primer On Subsurface Imaging
Stephen C. Myers, Lawrence Livermore National Laboratory

Different types of data — seismic, electrical, magnetic, gravity, radar, and geodesic — are sensitive to different physical properties in the subsurface, and the imaging objective of each study dictates applicable data types. Further, the scale of subsurface imaging ranges from the whole earth to the shallow structure beneath few square meters of topsoil. In seismic studies alone, signal sources, instrumentation, and methods used at various imaging scales are vastly different. For instance, global imaging relies on large, uncontrolled sources (e.g. earthquakes) and extremely sensitive instrumentation. Many sensor installations used for global imaging are remote facilities with state-of-the-art instruments that operate below the natural background “earth hum”. In contrast, imaging of basin structure typically utilizes controlled sources and an extensive, temporary network of simple sensors that are deployed expeditiously on the surface. The use of controlled sources eliminates a significant cause of uncertainty, but the networks are commonly deployed in noisy environments and redundant measurements are needed to clarify signals.

Tomography is used to project data measurements onto an image of the subsurface. Unlike medical tomography, tomography of the subsurface is inherently non-unique. Prior knowledge of basic earth structure and constraints from the physical properties of earth materials must be used to produce a realistic image.

This presentation will feature a case study that illustrates how combining data-science techniques with seismological constraints can refine a data set and bring previously unidentified features in earth’s mantle into focus. In this case a global tomographic data set is simultaneously reanalyzed to improve the accuracy and statistical characterization of seismic-phase travel times between earthquakes and a global network of sensors. The resulting model is improving routine location of seismic events and advancing basic science through the discovery of new features that are changing our understanding of earth’s history."

Imaging Fracture Networks Using Joint Seismic and Electrical Change Detection Techniques
Hunter Knox, Sandia National Laboratories

During the first phase of this SubTER project, Sandia National Labs (SNL) conducted a series of high-resolution crosshole seismic imaging campaigns designed to characterize induced fractures at Blue Canyon Dome, EMRTC, Socorro, NM. These fractures were emplaced using a novel energetic source, also designed at SNL, that limits damage to the borehole. Using a high-resolution tool, we were successful at resolving fractures with 3-inch vertical resolution over the 65-foot borehole during these campaigns. This case study, while applied at admittedly shallow depths and close borehole spacings, highlights the technical advances that allow us to isolate changes in seismic velocity and relate them back to individual stimulations and simultaneously achieve laboratory scale imaging resolution in a field setting (i.e. in situ). During the second phase of the project, which is currently underway, SNL and its collaborators (LBNL, LLNL, and PNNL) began developing emerging seismic and electrical geophysical imaging technologies that will characterize 1) the 3D extent and distribution of fractures stimulated from the energetic source, 2) 3D fluid transport within the stimulated fracture network through use of a particulate tracer, and 3) fracture attributes through advanced data analysis. The team has placed focus on advancing these technologies toward near real-time acquisition and processing in order to help provide the feedback mechanism necessary to understand and control fracture stimulation and fluid flow. This work is comprised of two primary tasks, the first of which is to collect a comprehensive set of 4D crosshole seismic and electrical data to image the fracture network generated from the aforementioned novel energetic source. In addition, autonomous seismic and electrical resistance tomography (ERT) data will be collected to image the migration of a tracer designed to enhance the electrical conductivity contrast of the fracture network. Near real-time 4D ERT imaging will be tested and demonstrated during this phase. The second task, which the team will begin work on this spring, is to use the data collected to 1) develop methods of estimating fracture attributes from seismic data, 2) develop methods of assimilating disparate and transient data sets to improve fracture network imaging resolution, and 3) advance capabilities for near real-time inversion of crosshole tomographic data. Although many challenges, both known and unknown, will need to be addressed before real-time feedback is possible in these situations, advancement of this technology will fundamentally change the way we engineer Enhanced Geothermal Systems, store CO2 at carbon sequestration sites, and access tight shale gases.

Joint work with Johnson, Timothy (PNNL); Ajo-Franklin, Jonathan (LBNL); Morris, Joseph (LLNL); Grubelich, Mark C. (SNL); Preston, Leiph (SNL); Knox, James M (SNL); King, Dennis K. (SNL).

A Multi-variate Geospatial Approach to Constrain Subsurface Properties and Reduce Uncertainty
Kelly Rose, US DOE, National Energy Technology Laboratory

Predictions of subsurface properties, such as depth, thickness, porosity, permeability, pressure and temperature, are important parameters for models and interpretations of subsurface engineered and natural systems. These studies are used to predict subsurface resource distributions (hydro, heat, hydrocarbon, mineral, storage capacity etc.), support environmental assessments, geohazard assessments, and provide a better understanding of geologic systems in general. However, the availability of data used to characterize these systems varies significantly, are generally acquired using indirect detection methods, and often are clustered resulting in areas with minimal or no data coverage. The traditional approach for evaluating and interpolating subsurface properties relies on either deterministic methods or stochastic methods. While there are some studies that utilize a combination of these methods to evaluate subsurface properties, the implementation of these methods remains inconsistent or incomplete. Thus, a robust, methodical and consistent approach for integrating geoscience knowledge with spatial statistical methods offers the opportunity to improve subsurface interpretations and reduce uncertainty about subsurface properties while also providing more information that can benefit advanced analytical models and analyses. In this presentation we will describe a hybrid deductive-probabilistic approach to support subsurface property characterization and uncertainty reduction. Specifically, describing how integration of deductive information about geologic systems, more traditionally associated with geoscience, can be paired with inferential methods associated with spatio-temporal quantitative analyses to improve certainty and further constrain subsurface properties.

Towards a High Resolution, Integrated Hydrology Model of North America: Diagnosis of Feedbacks between Groundwater and Land Energy Fluxes at Continental Scales.
Reed Maxwell, Colorado School of Mines

Recent studies demonstrate feedbacks between groundwater dynamics, overland flow, land surface and vegetation processes, and atmospheric boundary layer development that significantly affect local and regional climate across a range of climatic conditions. Furthermore, the type and distribution of vegetation cover alters land?atmosphere water and energy fluxes, as well as runoff generation and overland flow processes. These interactions can result in significant feedbacks on local and regional climate. In mountainous regions, recent research has shown that spatial and temporal variability in annual evapotranspiration, and thus water budgets, is strongly dependent on lateral groundwater flow; however, the full effects of these feedbacks across varied terrain (e.g. from plains to mountains) are not well understood. Here, we present a high-resolution, integrated hydrology model that covers much of continental North America and encompasses the Mississippi and Colorado watersheds. The model is run in a fully-transient manner at hourly temporal resolution incorporating fully-coupled land energy states and fluxes with integrated surface and subsurface hydrology. Connections are seen between hydrologic variables (such as water table depth) and land energy fluxes (such as latent heat) and spatial and temporal scaling is shown to span many orders of magnitude. Model results suggest that partitioning of plant transpiration to bare soil evaporation is a function of water table depth and later groundwater flow. Using these transient simulations as a proof of concept, we present a vision for future integrated simulation capabilities.

Joint work with Laura Condon, Syracuse University

Exascaling Nuclear Innovation
Rachel N. Slaybaugh, University of California, Berkeley

There is a growing need for innovation in nuclear detection, nuclear security and nonproliferation, and nuclear energy to meet our goals in global security, economic competitiveness, and environmental responsibility. There are incredible opportunities for progress in these important areas enabled by exascale computing and data analysis; however, we need to discover how to effectively capitalize on these opportunities. This work will touch upon three major challenges: performing computations on exascale architectures effectively and reproducibly, having sufficiently-accurate data for the highly-resolved solutions we need, and performing analysis on such large data sets efficiently.

Effective and reproducible exascale computing is essential for truly predictive simulation. Predictive simulation facilitates more rapid innovation than experiments alone for understanding detector materials and behaviors, stockpile stewardship, advanced reactor design, potential material diversion and false positives or negatives in monitoring, and other applications. We are developing algorithms to use traditional and advanced hybrid

architectures for neutral particle transport to enable predictive simulation at an affordable energy cost [1, 2]. We are also working with computer scientists to understand how to deal with the reproducibility challenges introduced by very large machines. Some aspects of these algorithms translate well to other applications.

The best algorithms, however, aren't enough if the data being used in the simulations is not sufficiently accurate. The ability to calculate highly resolved answers is illuminating challenges in creating and using cross section data sets as well as areas where better fundamental and covariance data is required [3, 4]. We are involved in efforts to modernize data formats and processing tools, strategically identify the highest impact data to improve first, and in experiments to generate the data itself.

Across nuclear applications there is an increasing need for quick, accurate answers: post-detonation response training, event response agility, design iteration, and inspection effectiveness. Conversely, increased resolution produces larger data sets that are increasingly difficult to move, process, visualize, and analyze at all, let alone in a timely fashion. One possible revolution to enable nuclear innovation is on the fly data processing and analysis. What if we built algorithms that managed and tracked data more strategically? What if analysis was built into the computation itself? What new strategies become possible? What are the tradeoffs?

Exascale computing and the datasets that come with it present amazing opportunities for nuclear innovation. Effective computer use, sufficiently accurate data, and strategic data analysis models are the keys to capitalizing on these opportunities and exascaling nuclear innovation.

[1] R.N. Slaybaugh, T.M. Evans, G.G. Davidson, and P.P.H. Wilson, "Rayleigh Quotient Iteration with a Multigrid in Energy Preconditioner for Massively Parallel Neutron Transport," Proceedings of Joint International Conference on Mathematics and Computation, Supercomputing in Nuclear Applications, and the Monte Carlo Method in Nashville, TN, April 2015.
[2] Ryan M. Bergmann, Jasmina L. Vujic. "Algorithmic choices in WARP — A framework for continuous energy Monte Carlo neutron transport in general 3D geometries on GPUs," Annals of Nuclear Energy. 77 (2015) 176-193.
[3] S.C. Wilson and R.N. Slaybaugh. "Improved Monte Carlo Variance Reduction for Space and Energy Self-Shielding," Nuclear Science and Engineering. 179 (2015) 22-41.
[4] Nuclear Data Needs and Capabilities for Applications Workshop, Department of Energy, Office of Science and Nuclear Science and Security Consortium, Lawrence Berkeley National Laboratory, May 27-29, 2015, https://bang.berkeley.edu/events/ndnca.

Partitioning a Large Simulation as It Runs

Kary Myers, Los Alamos National Laboratory

As computer simulations continue to grow in size and complexity, they present a particularly challenging class of big data problems. Many application areas are moving toward exascale computing systems, systems that perform a billion billion FLOPS (FLoating-point Operations Per Second). Simulations at this scale can generate output that exceeds both the storage capacity and the bandwidth available for transfer to storage, making post-processing and analysis challenging. One approach is to embed some analyses in the simulation while the simulation is running --- a strategy often called in situ analysis --- to reduce the need for transfer to storage. Another strategy is to save only a reduced set of time steps rather than the full simulation. Typically the selected time steps are evenly spaced, where the spacing can be defined by the budget for storage and transfer. Our work combines both of these ideas to introduce an online in situ method for identifying a reduced set of time steps of the simulation to save. Our approach significantly reduces the data transfer and storage requirements, and it provides improved fidelity to the simulation to facilitate post-processing and reconstruction. We illustrate the method using a computer simulation that supported NASA's 2009 Lunar Crater Observation and Sensing Satellite mission.

Local Gaussian Process Approximation for Large Computer Experiments
Robert B. Gramacy, The University of Chicago

We provide a new approach to approximate emulation of large computer experiments. By focusing expressly on desirable properties of the predictive equations, we derive a family of local sequential design schemes that dynamically define the support of a Gaussian process predictor based on a local subset of the data. We further derive expressions for fast sequential updating of all needed quantities as the local designs are built-up iteratively. Then we show how independent application of our local design strategy across the elements of a vast predictive grid facilitates a trivially parallel implementation. The end result is a global predictor able to take advantage of modern multicore architectures, GPUs, and cluster computing, while at the same time allowing for a non stationary modeling feature as a bonus. We demonstrate our method on examples utilizing designs sized in the tens of thousands to over a million data points. Comparisons are made to the method of compactly supported covariances, and we present applications to computer model calibration of a radiative shock and the calculation of satellite drag.

Leveraging Data Analysis to Improve Simulations
Ryan G. McClarren, Texas A&M University

The common modality for simulation-driven design of systems and experiments is to run a suite of calculations to optimize some set of parameters. Often the inputs on these simulations are small perturbations. Moreover, a given simulation will often have settings that, when chosen correctly, can cause the simulation to complete much faster. Inside a given design effort, knowledge from previous calculations is generally used by the modeler to optimize calculations, but cross-modeler and cross-project knowledge is not generally shared. Therefore, we propose to use the results of 'all' past simulations to inform both the direction of design (i.e., predict the result of a simulation before it is done), and inform the settings in the simulation as it runs. At exascale, such a data informed experiment will likely require in situ data analysis during a simulation. As a particular example I will discuss the application of data analysis to inform the mesh relaxation parameters in a simulation of an ICF experiment.

Joint work with Yuriy Ayzman, Texas A&M University