Workshop Presenter Abstracts:

Kevin Butler – Environmental Systems Research Institute
“Chicken or Egg: Using Machine Learning to Generate and Consume Large Collections of Ecological Data”

Our complex world is facing serious challenges on many fronts, from severe climate events to drought to food production challenges. Immediate action is necessary but as Richard Saul Wurman says, "understanding precedes action." Our science, infused with machine learning techniques, is foundational to that understanding. The availability of large spatial data sets, at increasingly finer spatial scales, requires that we use machine learning to both summarize massive collections of earth observations into *science information products* and to consume them to help solve the complex problems facing our planet. This talk will review three global, authoritative spatial data sets; Ecological Land Units (ELUs), Ecological Marine Units (EMUs) and a highly consistent and accurate global satellite-based elevation model which brings many improvements over SRTM 30. Ecological Land Units are a systematic division and classification of ecological and physiographic information about land surface features. Ecological Marine Units are baseline 3D mapped ecosystems of the ocean that have been classified through statistical clustering. The genesis, size, and quality of each data set as well as its applicability to machine learning techniques will be discussed.

Jennifer B. Dunn – Northwestern University
“Identifying and Quantifying Land Use Change through Machine Learning-Based Interpretation of Remote Sensing Products and Satellite Imagery”

After the Renewable Fuel Standard, a policy that set volumetric targets for various biofuels by target dates, was in place for nearly a decade, efforts began to characterize the LUC that could have been caused by agricultural expansion to produce biofuel, notably corn ethanol, feedstocks through remote sensing (RS) products. These efforts have improved with time, but still face challenges common to all efforts to quantify LUC with remote sensing products. These challenges include the need for novel and improved algorithms to classify RS data, for ecological insights to inform reasonable interpretation of LUC estimates from RS products, and for better and more consistent classification of uncertainty. Given these barriers, it is desirable to supplement RS-based interpretation of LUC with insights from high-resolution satellite imagery. Manual interpretation of such imagery is too time intensive to be implemented over a large scale. Machine-learning techniques are therefore being applied to interpret high-resolution satellite imagery and are expected to boost efforts to quantify changes in challenging land cover classes such as wetlands and grasslands. Given the large volumes of data to be interpreted to accomplish a national-level assessment of LUC at the resolution and confidence levels required to take steps to mitigate undesired LUC, machine-learning based approaches will be indispensable in the exploitation of RS products and satellite imagery.

Ian Foster – Argonne National Laboratory
“Introduction to Machine Learning”

This introductory talk is intended to establish a common foundation for subsequent discussion by defining significant terms and introducing important technologies.
**Bill Hargrove** – United States Forest Service  
“A Generic Imputer to Estimate Species Productivity After Future Global Tree Range Shifts”

I will describe a Generic Imputer that can produce a continuous gridded surface of imputed tree productivity values from a set of sparse point measurements, in this case the USDA Forest Service Forest Inventory Analysis plots across the conterminous United States. Values are imputed in a multivariate data space consisting of 17 individual descriptors of ecological growing conditions. While not a typical ML approach, the measurements can be viewed as "training," and the imputed values are not simply borrowed observations, but are adjusted appropriately. Both present and alternative future productivity surfaces are produced. By altering the interpolation methods that are used, the same Generic Imputer could produce estimates of other ecological variables like productivity, biomass, carbon, fuel, or flux exchange.

**Vipin Kumar** – University of Minnesota  
“Physics Guided Machine Learning: A New Paradigm for Modeling Dynamical Systems”  
Research funded by NSF and USGS

Physics-based models of dynamical systems are often used to study engineering and environmental systems. Despite their extensive use, these models have several well-known limitations due to incomplete or inaccurate representations of the physical processes being modeled. Given rapid data growth due to advances in sensor technologies, there is a tremendous opportunity to systematically advance modeling in these domains by using machine learning (ML) methods. However, capturing this opportunity is contingent on a paradigm shift in data-intensive scientific discovery since the “black box” use of ML often leads to serious false discoveries in scientific applications. Because the hypothesis space of scientific applications is often complex and exponentially large, an uninformed data-driven search can easily select a highly complex model that is neither generalizable nor physically interpretable, resulting in the discovery of spurious relationships, predictors, and patterns. This problem becomes worse when there is a scarcity of labeled samples, which is quite common in science and engineering domains.

This talk makes a case that in a real-world system that are governed by physical processes, there is an opportunity to take advantage of fundamental physical principles to inform the search of a physically meaningful and accurate ML model. Even though this will be illustrated in the context of modeling water temperature and quality in lakes, the paradigm has the potential to greatly advance the pace of discovery in a number of scientific and engineering disciplines where physics-based models are used, e.g., power engineering, climate science, weather forecasting, materials science, and biomedicine.

**Giri Prakash** – Oak Ridge National Laboratory  
“ARM Data Center: Next Generation data and computing architecture”

US Department of Energy’s Atmospheric Radiation Measurement (ARM) Data Center has been archiving and distributing various ground-based, aerial and model data products in support of atmospheric and climate research. The ADC Archive currently holds over 1.4 petabytes of data that dates back to 1992 containing over 11,000 data products. In support of DOE’s mission with an ever-increasing number of new field-campaigns and high-resolution modeling activities, ADC designed and developed modern and scalable architectures for processing, archival, analysis, discovery, and delivery.
In this presentation, the author will discuss about the ARM data catalog, end-to-end data lifecycle, data discovery and delivery, computing as a service using High Performance clusters. The presentation will also cover big data architecture using no-sql based ML platform.

Nicolas Tremblay – Agriculture and Agri-Food Canada
“Advancing Agriculture: ML and Geostatistics have their roles to play”

Agriculture is at the beginning of something new and it needs to achieve a better stewardship of resources at its disposal. Agricultural production is made in a complex spatio-temporal (and social) context. Old research methodologies have kept many away from geostatistical realities because spatial variability was seen for a long time as something to exclude rather than to integrate. The result has been that very few models have been so far able to grasp the key elements needed to optimize decision-making in order to balance economic and environmental requirements. Precision agriculture has not been performing up to expectations because of the inconsideration of temporal (mostly seasonal) features interacting with spatial patterns. Machine learning is emerging as a powerful opportunity to improve our success at predicting outcomes but its “black box” nature currently makes many uncomfortable. In any case, data and metadata have to be straightened up and made available to fulfill the calibration, validation and operational requirements of tomorrow’s farms.

Venkatram Vishwanath – Argonne National Laboratory
“Scientific Machine Learning and Data Management of the Argonne Leadership Computing Facility (ALCF)”

In this talk, I will first introduce the Argonne Leadership Computing Facility. I will next focus on facilitating scientific machine learning on supercomputing systems driven by application use-cases including, materials science, high energy physics and biosciences, as well the software stack to enable scaling on supercomputing systems. Finally, I will discuss various data management challenges and tools used by science teams on the ALCF systems.

Daniel Wieferich – United States Geological Survey
“ML Use Cases, Lessons Learned and Related Geospatial Data Assets”

Part 1: The USGS conducts a number of ongoing literature reviews that support science efforts by extracting "dark" information from literature into actionable databases. The development and maintenance of these databases often require intensive manual search and extraction methods. Through use cases of dam removal science and aquatic invasive species occurrence, we are exploring machine learning and text mining techniques to help keep these resources up-to-date in a timely, and less subjective manner.

Part 2: The USGS supports the development of a number of geospatial and natural resource data assets. This portion of the talk will give a brief overview of some USGS data products and projects that can be considered in future machine learning efforts in the natural resources field.