DuraMat Capability 1:
Data Management, Analytics and Informatics

Poster Number	Capability Name	Short Paragraph Description of Capability (300 words same as the abstract. Summarize what it does and value to DuraMat and Module Materials)	Capability Expert (principal contact)	Organization Name and Type - (National Laboratory - NL, Academic Institution - AI, Company- C)	Which Capability Area Best Fits This Work (Select One)	Define from an industry perspective what near term 1 year and long term 5 year successful use of the capability would be. (100 words)	Link to Your Website (if available)
1	DuraMat Capability 1: Data analytics for materials and module science	Many valuable data sets pertaining to solar materials and module performance remain scattered across institutions, are incompatible with one another, or are difficult to access and interact with. The Data Management and Analytics Thrust of DuraMat will build tools that unify solar materials data sets and data standards into a common platform while maintaining necessary access controls. This effort will also build links between data across length scales, e.g., between modules and their component materials. We expect this project to open new opportunities to use data analysis to understand degradation and failure patterns of solar PV modules. At LBNL, we have developed sophisticated frameworks for interacting with and disseminating large data sets to the community as well as in using fundamental materials property data to design new materials solutions. These efforts are embodied through the Materials Project effort, part of the Materials Genome Initiative, which currently hosts over 22,000+ registered users for exploring fundamental computational predicted data. We have also developed software packages based on industry standard tools such as Python, Pandas, Scikit-learn, and Plot.ly to interpret and analyze large data sets. We have further developed workflow packages (e.g., FireWorks) for large scale data analysis at large supercomputing centers. Thus a major software and data analysis capability exists at LBNL that can be applied to DuraMat data unification efforts. Proposed methods and ideas for analyzing solar PV data will be presented, and we look forward to discussing further opportunities for topics to explore with potential collaborators.	Anubhav Jain, LBNL (ajain@lbl.gov)	NL- Lawrence Berkeley	1. Data Management, Analytics, and Informatics	Year 1 - the ability to perform a unified analysis across multiple data sets that include fundamental materials data, module data, and external data sets (e.g., weather records) Year 5 - new models that explain and predict factors contributing to module failure based on their component materials and environment; a unified repository for data pertaining to solar PV	
2	Data2Design for Reliable PV Materials Performance	Sandia has a long history of developing advanced materials models, diagnostics, and characterization in order to assure the safety and reliability of our national security systems. We are interested in partnering with industry and academia to process and analyze existing & new data across different characteristic lengthscales leading to better and reliable materials and components selections. In order to transform the decision cycle with data utilization and analytics. We need to do the following. 1) Establish the necessary infrastructure to share existing data on materials performance and reliability that are of relevance to lifetime predictions of PV components. Analogous to the current infrastructure that Sandia has developed through its field collection and data harvesting techniques, this proposal focuses on materials reliability data. 2) Materials reliability and forensics data originate from materials characterization techniques that are inherently diverse. Characterization of materials aging or failure can range from microscopic images, spectra, finite-element meshes, and multidimensional text files. The granularity of experimental or simulated results is highly specific to the scale of analysis. Sandia needs an archival scheme that can be adapted to DuraMat search engines for failure analysis and lifetime assessment. 3) PV reliability data collected on the field currently are averaged over a large area which is difficult to link to part-level failure. Sandia is proposing to partner with others to identify and collect part-level or module-level reliability data that can identify points of failure that are traceable to materials performance.	Amy Sun acsun@sandia.gov & Bruce King bhking@sandia.gov	NL- Sandia National Laboratories	1. Data Management, Analytics, and Informatics	In the near term, this project will have portals in the DataHub that allow Sandia's legacy materials data be included and shared with our partners. In five years, a design-based tool is made available for accessing DuraMat database, and materials selection tied to specific performance goals.	
3	Data-driven Design of PV Module Materials: Informed by a Non-relational Data Warehouse and Analytics Environment with > 3.4 GW of PV Plant Datasets	The grand challenges that DuraMat faces in achieving a step change improvement in PV Module materials' combined cost, performance and lifetime fall into two areas: 1. How to dramatically increase the foundational data used in predictive materials design and enable new data science driven, degradation science studies [French, 2015]. 2. How to harness the broad experiments and simulations done by PV research community, so the aggregated research guides the development of DuraMat materials. The answer to these two challenges lies in A) bringing operating PV plants into the research enterprise as a critical epidemiological population for study, and B) using data science methods, such as machine learning and predictive network modeling, to aggregate and integrate DuraMat research results into common systems-level models of PV modules exposed to real-world conditions and lab-based accelerated exposures. Incorporating the 180 GW of real-world PV power plants into DuraMat research requires Petabyte scale distributed computing approaches combined with non-relational data warehouse methods. We have implemented [Hu, 2016] the Energy-CRADLE system and ingested time-series power, weather and insolation data from > 3.4 GW from 780 field deployed power plant sites with 5500 inverters across 13 Köppen-Geiger climatic zones spanning up to 15 years of operation. Energy-CRADLE enabled us to study 2.2 million I-V curves and identify the occurrence of 'step' I-V curves associated with increasing module outdoor exposure times. [Peshek, 2016] We are also ingesting stepwise point, spectral, image and hyperspectral image data from our PV degradation science studies [French, 2015], to enable cross-correlation of indoor lab and outdoor real-world exposures studies. Energy-CRADLE represents an operational petabyte and petaflop computing system, as imagined in the National Strategic Computing Initiative [Obama, 2015], and can be made accessible to the DuraMat consortia to provide an additional foundation for data-driven DuraMat materials design.	Prof. Roger H. French	AI- Case Western Reserve University	1. Data Management, Analytics, and Informatics	1 year: Bring operating PV plant time-series data into DuraMat research to, for example, define degradation rates of PV modules in different climatic zones, for Modules produced over the past 15 years. 5 years: Compare new DuraMat Materials lab and real-world performance in field deployed systems in RTC, Solar Lifetime, and commercial power plants.	http://sdle.case.edu
4	Building a Data Hub to Enable an Informatics-Driven PV Durability Materials Research	Materials research and development has become an increasingly data intensive endeavor over the past two decades. The ability to securely and effectively gather, marshal, aggregate, process, and share data has often been a deciding factor in the success of many projects and with cross–institutional consortia it is imperative. Beyond being a simple archive and communication platform, the DuraMat Data Hub must allow for the easy upload and download of PV durability related raw data products, analyzed data, modeling and simulation data, and data streams from deployed historical databases. By building on our previous successful experience with the Laboratory Informational Management System (LIMS) deployed at NREL, we will automate these processes where we can and where that is not practical, manual interfaces will be provided. Additionally the DuraMat Data Hub will also provide a means to take advantage of informatics methods and analysis for exploring the PV durability related data; therefore standards for clean and complete data formats must be developed, implemented, and supported within the archive and application layer interface (API) of the Hub. Due to the nature of research, security will be an important component to the Data Hub and it must provide project-level security to protect data that is deemed intellectual property or embargoed as well as provide a method to stream data into public repositories once that data has been considered releasable. With these criteria in mind, we are designing and implementing a data hub for the consortium where all members and partners will be able to archive, track, explore, and share data securely, efficiently, and effectively.	Robert White and Kris Munch	NL- NREL	1. Data Management, Analytics, and Informatics	During the first year we will be able to successfully begin to archive, protect, search and access data supporting the DuraMat consortium. By the five year point we will be continuing the curation of the archived data along with being able to cross-correlate data from variety of projects and data streams and seamlessly apply informatics techniques to the data to quickly and efficiently. We will be able to process data for analysis, locate trends and behaviors, and discover non-trivial parameter spaces, potentially leading to advanced discoveries.	
5	The NREL High Throughput Experiments for Materials (HTEM) Database - A Prototype Project-Specific Analytics Database Enabling the Application of Machine Learning to Experimental Data	To effectively use high-throughput experiments (HTE) as envisioned within the Energy Materials Network requires not only rapid experiments and automated data harvesting but also expression of the data and metadata in an accessible database to enable analysis and mining. Using the challenge of rapid analysis of large amounts of x-ray diffraction data arising from mapping of composition gradient thin film combinatorial libraries, we have created a prototype project-specific analytics database, the High Throughput Experiments for Materials database (HTEM DB). Project relevant experimental data and metadata harvested by the NREL Laboratory Information Management System (LIMS) is automatically extracted, pre-processed when appropriate and then expressed in the HTEM DB which currently contains x-ray diffraction, compositional, electrical and synthesis process data for ~ 50,000 effectively different samples. This data can be accessed via an interactive web interface, an applications programming interface (API) or a structured query language (SQL) interface. Project specific data visualization and machine learning based analytics tools have been implemented in a custom package, which is built upon the substantial open source machine learning resources available for Python. This analytics package can be run from the command line, from a Python notebook or from within any data analysis software capable of issuing a system command. Using this last approach, we have implemented a custom user-friendly analysis tool in Igor Pro, a commercial analysis program widely used at NREL. The overall result is a scientist-friendly extensible analysis environment with project specific machine learning and visualizations. Using the infrastructure of the existing HTEM DB as a prototype, project specific analytics databases and tools could be easily developed where desired for DuraMat projects.	John Perkins, Andriy Zakutayev, Caleb Phillips, Jacob Hinkle, Robert White, Kristin Munch, Marcus Schwarting	NL- NREL	1. Data Management, Analytics, and Informatics	Project specific analytics databases and the associated user tools can provide project participants with easy access to project data as well as visualization and/or analysis tools customized to the needs of the project. By templating off of existing prototypes, such project specific databases could be created and put into use in the near term (1 Year). In the longer term (5 Years), the data they would contain as well as the project focused analysis tools developed would become increasingly useful.	
6	Data Analytics for Mining Process-Structure-Property Linkages in Hierarchical Materials	A majority of the materials employed in advanced technologies exhibit hierarchical internal structures with rich details at multiple length and/or structure scales (spanning from atomic to macroscale). Although the core connections between the material's structure, its evolution through various manufacturing processes, and its macroscale properties (or performance characteristics) in service are widely acknowledged to exist, establishing this fundamental knowledge base has proven effort-intensive, slow, and very expensive for most material systems being explored for advanced technology applications. It is anticipated that the multi-functional performance characteristics of a material are likely to be controlled by a relatively small number of salient features in its hierarchical internal structure. However, cost-effective validated protocols do not yet exist for fast identification of these salient features and establishment of the desired core knowledge needed for the accelerated design, manufacture and deployment of new materials in advanced technologies. The main impediment arises from lack of a broadly accepted framework for a rigorous quantification of the material's structure, and objective (automated) identification of the salient features that control the properties of interest. Our ongoing research is focused on the development of data science algorithms and computationally efficient protocols capable of mining the essential linkages from available ensembles of materials datasets (both experimental and modeling), and building robust knowledge systems that can be readily accessed, searched, and shared by the broader community. The methods employed in this novel framework are based on digital representation of material's hierarchical internal structure, rigorous quantification of the material structure using n-point spatial correlations, objective (data-driven) dimensionality reduction of the material structure representation using data science approaches (e.g., principal component analyses), and formulation of reliable and robust process-structure-property linkages using various regression techniques.	Prof. Surya Kalidindi, surya.kalidindi@me.gatech.edu, & Hojun Lim, hnlim@sandia.gov, Sandia National Laboratories	AI- Georgia Institute of Technology NL- Sandia National Laboratories	1. Data Management, Analytics, and Informatics	In the near term, this project will help establish essential tools for the DuraMat DataHUB. In five years, tools developed will accelerate new PV materials identification based on desired or measured materials microstructures that meet performance targets.	