CIF21 DIBBs: EI: Data Laboratory for Materials Engineering (Award # 1640867)

Venu Govindaraju (PI), Krishna Rajan, Thomas Furlani, Srirangaraj Setlur, Scott Broderick

This project is integrating the building blocks of document processing and machine learning algorithms, which we are applying to materials science problems. From these building blocks, we will accelerate the rate of discovery beyond what is possible through a trial-and-error search strategy. This project directly addresses the goals of the Materials Genome Initiative (MGI) to accelerate the pace of discovery and deployment of advanced material systems. To obtain insights for the discovery of new materials and to study existing materials, R&D scientists and engineers rely heavily on an ever-growing number of materials research databases and publications, mostly available online, that date back many decades. So, the major thrust of this research work involves using technology to (i) extract deep meaning from a large corpus of relevant materials documents and data from scientific databases; (ii) navigate, cluster and present data in a meaningful way; (iii) evaluate and revise the materials-related query responses until the researchers are guided to their information destination and (iv) facilitate faster discovery of new materials using the enhanced knowledge base. While this methodology targets the interdisciplinary field of materials research, the tools to be developed can be generalized to enhance scientific discoveries and learning across a broad swath of disciplines.

The infrastructure building blocks will enable researchers in a variety of scientific domains to utilize the power of machine learning for extracting rich information from scientific publications, technical handbooks and databases that can facilitate innovative data analytics approaches. The building blocks include (i) a document image processing toolkit for deriving insights from information-rich constructs such as graphs and plots; (ii) a machine learning toolkit for materials discovery and data analytics; and (iii) a visualization toolkit for information retrieval and insightful analysis. The machine learning toolkit will feature innovative application of algorithms such as hierarchical, dynamic topic models to investigate trends in materials discovery over user-specified time periods. The field of image based document analysis will benefit tremendously from the use of tools such as deep belief networks for classification and analysis of information-rich constructs such as phase diagrams and tables from research papers as well as materials databases. Developing interactive visualization tools that can model large materials networks from multiple perspectives will lead to an advancement in visualization studies.

The data analytics framework integrates various data formats with machine learning tools for an accelerated design procedure. This further builds descriptor databases, as opposed to the standard approach of building property databases. Still the current design process is slowed by manual integration of data information due to the variety of data forms and data complexity. The current process is both too slow and introduces too much user bias in the procedure. This project addresses these challenges by developing a framework which both standardizes and accelerates the process. A description of the anticipated platform is as follows. The researcher interacts with our engine via an interactive questions/answering session at the front-end. An extensive documents-driven, indexing system is pre-built at the back-end from tables, figures, range of data, etc. Data in the engine is organized primarily by materials, material-related phenomena and by key properties of the materials. There is also a time component to all the data. Based on the nature of the query, the system can either access the pre-built hierarchical statistical model or dynamically develop a statistical model for the specific request. The results are presented to the user via the interactive visualization tool to browse over time, and over related materials. To perform direct query searches using material properties and data ranges, results can be displayed as tables, figures and plots from the articles. The core building blocks such as (i) the hierarchical, time-based, clustering statistical model, (ii) the ability to search on texts in plots and tables and display the query results in form of the original figures, and (iii) the interactive visualization tool developed under this proposal will help us achieve this vision.