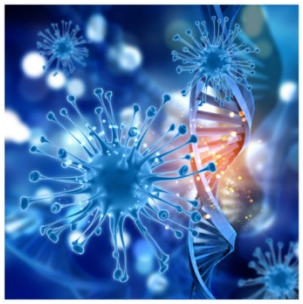
# A shared platform for big data and high-performance computing that supports the entire research process end-to-end



We offer the Digital Environment for Enabling Data-Driven Science (DEEDS) to scientific and engineering communities everywhere, as a full-service platform that provides end-to-end support for your research investigations.

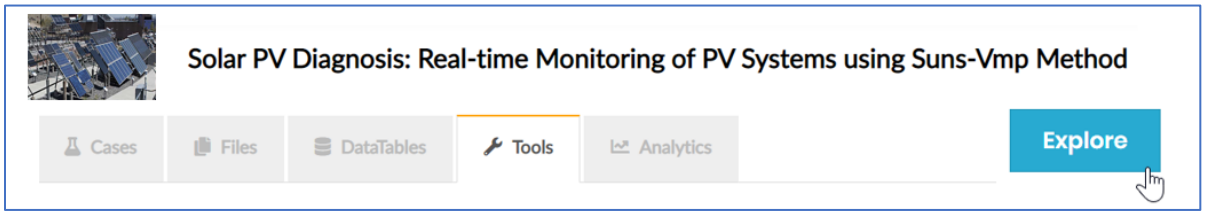
DEEDS is a systematic, secure, reliable way to conduct your research, with powerful yet user-friendly interactive support for big data, high-performance computing and shared scientific workflows.

The design and development of DEEDS is a collaborative effort, bringing together scientists and engineers from chemistry (molecular dynamics, quantum chemistry), agriculture (environmental science, ecotoxicoloy, forest biodiversity, microbiome engineering), electrical engineering (solarPV efficiency diagnosis), civil engineering (bridge health monitoring, earthquake engineering), health and human services (nutrition science, clinical studies), biology (RNA sequencing) and computer science (infrastructure, big data, high performance computing). The platform we developed supports research activities throughout the research investigation and is effective across science domains.

## Why DEEDS?

In research investigations, scientists and engineers carry out complex, long-running, team-shared activities involving the collection and integration of data, the execution of computational software, the analysis and visualization of data and software output, and the synthesis of results to produce conclusions. In most disciplines this process is done in an ad-hoc manner because existing IT platforms support only part of the investigative workflow. Many ongoing research projects do not use cyberinfrastructure to support any part of their shared investigations. This means that data, computing, analysis, and outcomes reside in disconnected environments. The lack of continuity in the research process hinders it in many ways: the process is difficult to share and difficult to validate. It compromises reproducibility of results and makes reuse and reinterpretation of data and algorithms more difficult.

**With DEEDS ... research data, computing, and scientific workflows come together in datasets that you build, use and share across your entire investigation.**




Research teams build and share DEEDS datasets using an interactive **dashboard**.

The dashboard provides full-featured services for the organization and structuring of research activities; the upload and classification of data; the extraction and assignment of metadata; research computing and statistical modeling (including HPC computing); the automatic capture of scientific workflows for data provenance and reproducibility; and the analysis and visualization of results.

User-friendly interfaces help research teams create, annotate, and link file repositories, multi-dimensional, hierarchical data tables, computational software, statistical models, scientific workflows and analytics – offering interactive search, exploration, and visualization across all elements of the dataset. Datasets are FAIR-compliant and can be published for discovery and interactive exploration of data, computing tools, and workflows for reuse and reinterpretation.

The DEEDS dashboard provides services for

- **Data**: Upload, preserve, manage and explore your data. Assign metadata, follow rules for metadata standards (exceeding FAIR compliance), integrate scripts to automatically transform, validate, curate and check completeness of your uploaded data. Dataset data consist of

    - **Files** classified by type, format and use, including standard categories and user defined categories (e.g., sensor data, mass spectrometry data, geospatial data, protocols). Files can be imported from external repositories (e.g., DuraMat through CKAN API). DEEDS offers applications to search, explore and visualize data by type (e.g., geotiff tile generation for map overlays). Some DEEDS dataset repositories have more than 3M files that are indexed, linked and classified by DEEDS for fast, user-friendly navigation, search and visualization.

    - **Data Tables** that represent hierarchical, multi-dimensional data models for measurements, properties, observations and other data. These can be customized, organized, re-organized, cloned, and annotated across the investigation lifecycle. Users can upload spreadsheets or interactively update (including bulk updates). Data tables can be viewed, browsed, searched, filtered, and downloaded. Data table operations are robust and user friendly. Some DEEDS data tables have more than 300 columns, some have more than 11M rows. Data table cells can represent a single data point, data arrays, and single/multiple linked tabular datasheets. Data table updates, viewing, search, filtering and exploration remain robust across all representations. Data tables are linked to map overlays for unified geospatial exploration, and data tables are the fundamental analytics structure for visualization, computing and analysis.

- **Computing tools**: Define computing tools to your dataset, then launch and track execution workflows. Your dataset computing tools can be computational research codes, open source software packages, modeling scripts, licensed software, Jupyter notebooks, RShiny and other Hubzero tools. DEEDS tools have full access to your dataset repository. A tool launched from the DEEDS dashboard follows the owner's tool definition, allowing users to specify tool arguments, choose input files, select execution resources (including HPC facilities), and determine how the computing workflow should be tracked and captured. Output data are returned by DEEDS to the dataset – they are annotated and linked to input, tool, resource and user. Captured workflows can be viewed and searched for data provenance and results traceability. Tools defined for DEEDS

datasets are added to the DEEDS tool repository, where they can be imported into other datasets with permission of their owners.

- **Analytics**: Define R data frames based on dataset files and data tables. DEEDS R-based analytics supports data filtering, merging, statistical computing, algorithm specification/ computation, and visualization. Data across datasets as well as external data can be merged within analytics for comparison and analysis. Introduction of new analytics features is ongoing, based on priorities of DEEDS research groups.
- **FAIR compliance**: DEEDS guarantees adherence to the principles of Findable, Accessible, Interoperable, and Reusable data management and stewardship for your research. DEEDS offers fine-grained access control as needed for your data and computing tools.

## Funding

DEEDS research is supported by the National Science Foundation (NSF) under Grant No. 1724728, CIF21 DIBBs: EI: Creating a Digital Environment for Enabling Data-Driven Science (DEEDS), awarded by the Office of Advanced Cyberinfrastructure (OAC), Directorate for Computer & Information Science & Engineering. NSF awards $3.5 million 4-year grant to build powerful web platform for data-driven science. Data cyberinfrastructure work for DEEDS supported by previous NSF AOC award No. 1443027.

## Contact Us

**Principal Investigator, Co-investigators & Senior Personnel**
Ann Christine Catlin acc@purdue.edu
Muhammad Ashraful Alam alam@purdue.edu
Marisol Sepúlveda mssepulv@purdue.ed
Joseph Francisco frjoseph@sas.upenn.edu
Kathleen Hill Gallant hillgallant@purdue.edu
Chandima Hewa Nadungodage chewanad@purdue.edu