BIG DATA ILLUMINATES the Physical Sciences

LAWRENCE Livermore's Physical and Life Sciences (PLS) Directorate is teeming with data. In research areas ranging from biomedicine and nuclear chemistry to high-energy-density physics and climate science, data from experiments and simulations accumulate quickly as ever more sophisticated tools are developed to collect it. Although manual data analysis may have been manageable in the past, this newfound abundance of information requires state-of-the-art, often automated techniques to achieve reliable interpretation.

Similar to their peers across the Laboratory, many PLS researchers are turning to data science to address Livermore's unique application-specific challenges. Data science includes many related disciplines such as artificial intelligence, big-data analytics, computer vision, machine learning (ML), predictive modeling, statistical inference, and uncertainty quantification. (See *S&TR*, March 2019, pp. 4–11.) Fittingly, this rapidly growing field has become a focus for projects funded by the Laboratory Directed Research and Development (LDRD) Program. For fiscal year 2019, two PLS principal investigators—physicist Michael Schneider and materials scientist Yong Han—were awarded LDRD Strategic Initiative (SI) projects that apply data science to relevant PLS research. LDRD-SI projects are large in scope and address key science, technology, and engineering challenges related to Livermore's strategic planning.

For these initiatives, Schneider and Han have built multidisciplinary teams in which domain scientists work alongside data scientists, applying data analysis and interpretation techniques to inform scientific exploration. "Data is a tremendously

valuable commodity and just as important as physical inventory," says Han, who recommends that for these types of projects, investigators devise a strategy to capture the right data, manage it, and formulate a legacy plan around it. Schneider adds, "Finding the expertise and resources to integrate data-driven approaches into the project from the beginning is an essential first step."

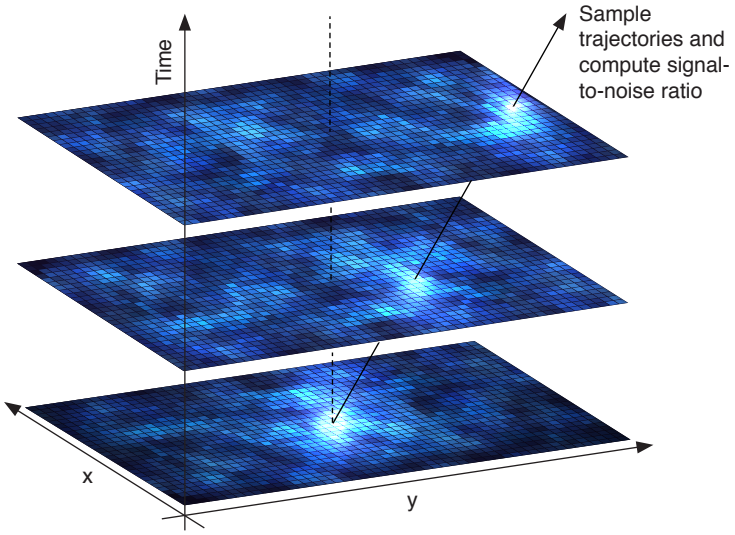
Mysteries of the Universe

Astrophysics is a growth area in the Laboratory's advancement of basic science for national and global security needs. In this field, data science helps researchers catalog and interpret objects orbiting Earth and process huge volumes of data captured by ground- and space-based telescopes. Schneider says, "We are closing gaps in our understanding of how the universe works. Astrophysics data holds clues to resolving some of the most pressing unanswered questions."

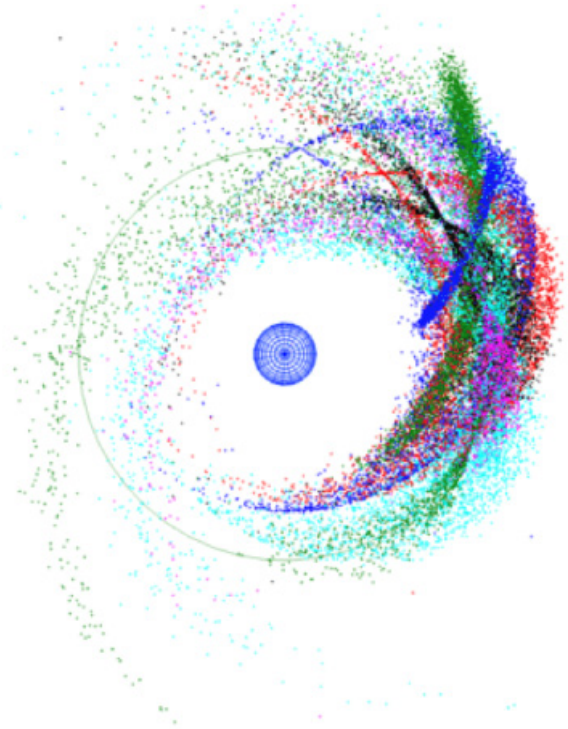
In an LDRD project that began in 2016, Schneider and colleagues studied the nature of dark energy—matter of unknown composition that does not emit or reflect electromagnetic radiation and is therefore difficult to observe directly. Specifically, the team developed image-analysis algorithms that make inferences from low signal-to-noise data. This process requires evaluating statistical versus systematic errors in the data set. Whereas statistical errors shrink as the data set grows, systematic errors are caused by variables that do not simply average out—for instance, atmospheric turbulence or imperfections in a camera lens. Gravitational lensing, wherein light from a distant object is bent, and thus distorted, by a massive object in the field of view, is a signal the team looks for in observations of galaxies and other light sources. Schneider's team accounts for these errors with probability-based Bayesian models, simulations of noise, and data-processing software. "Measuring the average energy density of empty space in the universe is complicated because everything else gets in the way. We're re-envisioning a data-processing pipeline to evaluate and correct those errors," explains Schneider.

Similarly, another PLS-led LDRD project seeks to identify the gravitational lensing signatures of black holes. The research team, led by principal investigator Will Dawson, is processing 12 terabytes of imaging data—for more than 500 million stars—collected from telescope surveys. By combining observations and simulations of light curves, the team aims to make the first direct measurement of black hole mass spectra in the Milky Way galaxy. The ability to analyze such large volumes of data is also valuable for studies of other astrophysical phenomena, such as asteroid trajectories.

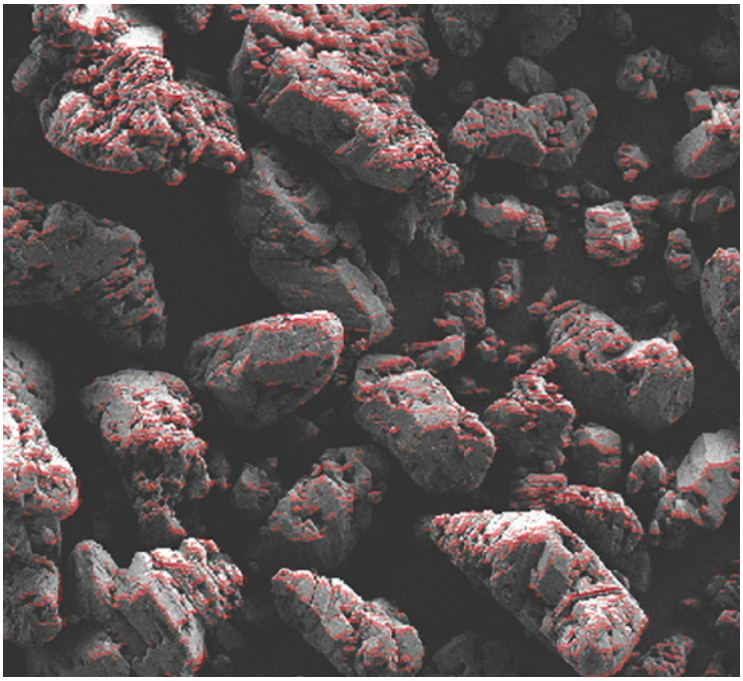
Schneider's LDRD-SI project combines Bayesian statistical methods and deep neural networks (a subset of ML) to analyze debris, satellites, and other objects orbiting Earth. Unlike the dark energy and dark matter studies, these telescope-generated data sets are comparatively small. "We need a newer predictive physics model to achieve higher accuracy using the available



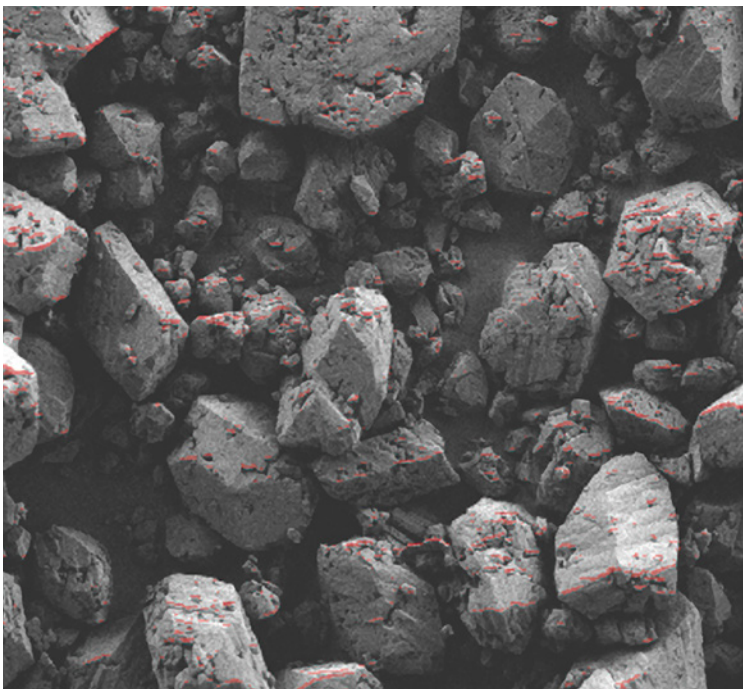
A Livermore-developed image-processing algorithm performs digital frame stacking to detect a moving asteroid in images captured by the Department of Energy–funded Dark Energy Camera, located at the Cerro Tololo Inter-American Observatory in Chile.



Data analytics algorithms generate probabilistic tracks of Earth-orbiting objects that are seen for only short time periods. The particle "clouds" shown here (represented by different colors) indicate the candidate positions for each object as projected into the future. (Blue circle represents Earth.)



Computer vision techniques are automated tasks that enable computers to analyze digital images and highlight specific regions of interest. In these scanning electron microscopy images, researchers are looking for ML-predicted features (shown in red) in high-performance (top) and low-performance (bottom) materials.



data," notes Schneider, whose team has developed ML algorithms that predict orbital paths in this data-starved environment.

Looking ahead, similar techniques could enable collaborative autonomy within a constellation of satellites tasked with tracking orbiting objects. In this setup, each satellite exchanges data with the rest of the network until they reach consensus on an object's location. This capability could help predict when space debris will hit a satellite or Earth. Schneider says, "We can make the most of big data using practical computing and by extracting interpretable features using data science techniques."

Optimized Materials

Lawrence Livermore's innovative materials include advanced metallic nanowires, high-performance alloys, freestanding polymer films, components that refresh the aging nuclear stockpile, and more. Regardless of the final product, feedstock materials (those raw materials directly used in the manufacture of other products) must be synthesized and optimized for system-level integration that will meet performance requirements with predictive behaviors.

Data science techniques are at the heart of projects focused on accelerating materials discovery, optimization, and deployment processes. Typically, the materials discovery process takes 10 to 15 years before application integration. "To improve the development cycle, we should leverage new tools, especially at the Laboratory where people across all disciplines are working on data analysis problems and solutions," states Han, who leads several of Livermore's data science–supported materials development projects.

One LDRD project was inspired by the growing volume of scientific literature. Materials scientists publish tens of thousands of papers every year that contain valuable information about the "recipes" they used to generate new materials. Han says, "Experimentalists must read a great deal of literature to stay informed in their fields while learning the new protocols and chemicals others have used in creating materials." His team created a browser-based tool to extract targeted information from the published papers, thus automating a repetitive, formidable task. (See *S&TR*, July/August 2017, pp. 16–19.) The extraction pipeline begins with a supervised logistic regression algorithm that highlights recipe-like text sentences. Another algorithm combines a conditional random-field model with natural-language processing to extract chemical information—formulas, concentrations, relationships, reaction times, and morphology, among other variables. Visualizations render the data for further analysis by end users.

Another project, sponsored by Livermore's Weapons and Complex Integration Principal Directorate and supported by the Weapon Simulation and Computing Program, goes beyond text descriptions of materials by analyzing images of high-explosive (HE) materials, whose physical properties often correlate with performance but are difficult to quantify. Han's research team developed a feature extraction tool that determines

At the Data Science Institute's 2018 workshop, Livermore staff scientist Yong Han described a project that uses ML algorithms to extract targeted information from scientific papers emphasizing nanomaterials synthesis. Efficiently gathering data from the literature and rapidly analyzing complex relationships will accelerate the materials discovery, optimization, and deployment processes. (Photo by Ian Fabre.)

which features—among the hundreds revealed in a single scanning electron microscopy image—are meaningful. The tool uses open-source technologies that define engineered features, such as boundary detection, at a pixel level. The computer learns to weigh feature importance, then provides computed values that translate to mechanical performance prediction.

Han's LDRD-SI project further enhances the development of HE materials by combining multimodal data for feature extraction. With the help of data visualizations manipulated in a custom-built user interface, the team will correlate additional material properties with HE performance by identifying features in images and numerical values from varied sources. The team aims to advance the application of ML algorithms for small data sets while also implementing physics-based approaches. Robert Maxwell, leader of the Laboratory's Materials Science Division, states, "Our materials researchers generate terabytes of data on an hourly basis. Data science advances our ability to understand, develop, and deploy new materials. Tools that unite our multidisciplinary teams are the most powerful of all."

Building a Community

A collaborative data science community is a vital part of Livermore's investments in mission-driven research. The Laboratory's Data Science Institute (DSI) has quickly become a useful resource for researchers seeking expert analysis outside their scientific fields. DSI promotes multidisciplinary collaboration, enabling researchers to strengthen their work by applying data science expertise. Schneider, a member of DSI's governing council, states, "Our data science capabilities are as fundamental to Livermore's missions as our high-performance computing resources."

Scientists may not know how best to integrate data science techniques into their research, so DSI organizes on-site educational activities, such as reading groups and seminars. The DSI Consulting Service advises researchers on experimental design, data collection and sampling, and data-driven solutions for their projects. DSI's outreach extends beyond the Laboratory with technical workshops, invited speakers, and student internships. Such community-focused resources are increasingly valuable tools for 21st-century researchers.

With the help of PLS collaborators, the Laboratory's data science community is doing more than providing expertise on mission-driven projects. The field itself is expanding, with Livermore contributing to theoretical research in areas such as sample optimization and ML model interpretability. Schneider adds, "We have a responsibility to conduct quality science and produce reliable results, which requires innovation. The Laboratory has a leg up in advancing data science techniques because we have the sophisticated scientific expertise to understand and apply the underlying math."

—Holly Auten

Key Words: algorithm, astrophysics, big data, computer vision, dark energy, dark matter, data science, Data Science Institute (DSI), high explosive (HE), Laboratory Directed Research and Development (LDRD) Program, machine learning (ML), materials science, Physical and Life Sciences (PLS) Directorate, statistics.

For further information contact Michael Schneider (925) 422-4287 (schneider42@llnl.gov) or Yong Han (925) 423-9722 (han5@llnl.gov).