S&TR March 2019

MACHINE LEARNING ON A MISSION

Livermore computer scientists advance machine learning technology for scientific applications.

As valuable tools for analyzing data from scientific simulations and experiments, machine learning (ML) algorithms are run on many of Livermore's high-performance computing resources, such as the new Sierra supercomputer. (Photo by Randy Wong.)

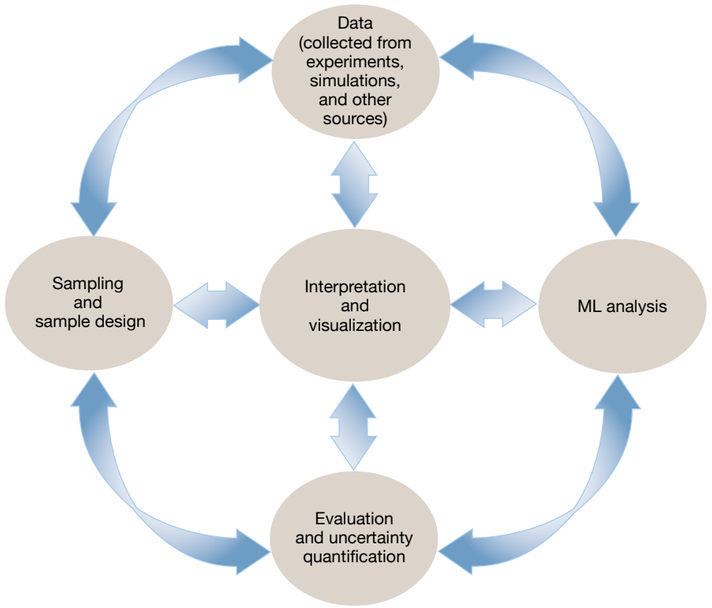
MORE than just a buzzword, machine learning (ML) has become part of everyday life. Social media platforms recognize faces in photos. Online stores recommend products related to shoppers' browsing and purchasing behavior. Smartphones offer word-completion suggestions based on users' texting habits. Search engines refine results after learning from users' past actions. Only with ML technology can self-driving cars adapt to moving traffic.

ML uses computers to learn from data and make predictions about the environment. As the world generates more data, interpretation becomes more difficult. Lawrence Livermore computer scientist Peer-Timo Bremer explains, "Humans reach a limit where they cannot perform the analysis anymore."

A smart machine—one that adapts to new information on the fly—can speed up processing and analysis times and improve its accuracy in identification and prediction tasks. Although commercial and consumer applications of ML are numerous, Livermore's mission space also presents ample opportunities for exploiting ML tools, often requiring new development beyond standard applications. (See box p. 7.)

Indeed, Livermore faces unique challenges in advancing the ML arena. Bremer points out, "Commercial companies do not solve scientific problems, just as national laboratories do not optimize selections of movie reviews. We therefore build on commercial tools to create the techniques we need to analyze data from experiments, simulations, and other sources." ML algorithms must be scaled for high-performance computing (HPC) machines, and different types and varying volumes of data complicate matters. For example, one project may have access to thousands of patient health records, whereas another may only have data from a handful of National Ignition Facility (NIF) shots. Bremer continues, "A team may have to sort through genetic sequences, protein structures, energy spectra, x-ray images, or combinations of these." Other issues with scientific data include noise and imbalance—such as a handful of successful drugs versus millions of ineffective compounds—which will bias traditional data-driven models.

Along with Bremer, computer scientists Rushil Anirudh, Harsh Bhatia, Bhavya Kailkhura, Hyojin Kim, Shusen Liu, and Jayaraman Thiagarajan are go-to ML experts. They take a bidirectional approach, both advancing underlying theory and solving real-world problems. The algorithms involved are run on several on-site HPC resources, including Sierra, the Laboratory's newest and fastest supercomputer.



In scientific analysis using ML, sample design informs data collection in simulations and experiments. The ML model processes the data and generates predictions, which are then evaluated for quality. Results from both training and actual data are fed back into the sample design to refine the process. Visualizations are used for interpretation.

Perfecting the Process

Scientific analysis involving ML generally follows a cycle in which sample design guides data collection. Data are processed with ML algorithms and their associated frameworks—collectively, the ML model—which are designed to learn from data inputs. Results are scrutinized for errors and unknown variables, providing statistical quantification of uncertainties and informing subsequent sampling. All stages of the cycle are interpreted with visualization tools. The ML model is first trained on smaller, representative data sets to refine this process.

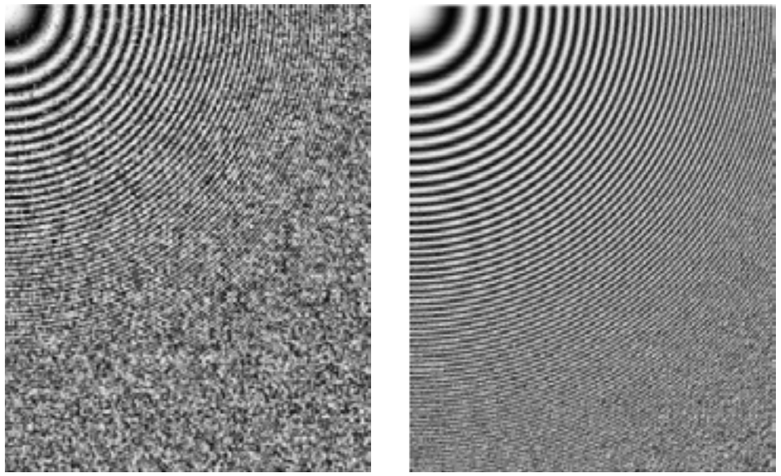
ML algorithms serve various purposes. For instance, neural networks (NNs) connect artificial neural units to observe and make inferences from data. Deep learning is another category of algorithms in which hierarchical layers of NNs adaptively learn from data to discover new features. In addition, ML methods respond differently depending on data properties. In supervised ML, the system analyzes labeled or classified data. In unsupervised ML, data are not labeled or classified, so the computer learns to identify common traits. Other types of learning are self-supervised—labeled and unlabeled data combined—and reinforcement—based on prior performance.

Livermore researchers actively develop new ways of configuring and deploying such algorithms. The common thread is improving ML's accuracy and efficiency for the benefit of the entire scientific analysis workflow. Accordingly, Thiagarajan explains, "All application domains face the same issue, and the conversation must start with the kind of data needed. Scientific analysis is driven by data."

Designing the Data Sample

Sample design is the key to quality results, especially when a project faces scope or resource constraints. Kailkhura notes, "Sampling requires configuring experiments and simulations to generate the most informative data. ML algorithms will not offer new insight if samples contain inaccurate or incomplete information." For instance, if a team with limited computing resources wants to run simulations of NIF shots, beforehand they must choose, as the focus of their investigation, the most valuable parameters. "These are the parameters," Kailkhura explains, "that will acquire the most information given a number of simulation runs, such as implosion dimensions."

Kailkhura looks at sample design abstractly, seeking mathematical solutions for sampling optimization problems. A high-dimensional (HD) parameter space is needed to represent the key factors that affect the results of a complex experiment or simulation. The higher the dimensionality, the more data are required to sample the space. Kailkhura describes these spaces from a theoretical perspective, citing the widely known sphere-packing problem—finding the ideal arrangement of oranges in a crate for n dimensions. In this problem, oranges represent data points in a sample, and the crate is the domain of interest, as in an inertial confinement fusion (ICF) implosion. The way the oranges are packed signifies the pattern of selected data points. Optimized sphere packing, or space filling, enables ML models to process data more quickly by minimizing the number of steps to reach a solution. Moreover, the models can provide insights into data not acquired yet, hence ML's predictive capabilities. Kailkhura seeks to cover as much of the space as possible while also obtaining the greatest information from the data sample. He states, "We strive for the right balance between coverage from uniform sampling and maximized information from random sampling. The optimal sample design will have some combination of uniformness and randomness."



(left) Random sampling of image reconstruction data finds only one significant pattern, as shown by concentric rings (upper left corner). (right) The team's spectral-sampling method reduces noise and other artifacts to reveal additional zones of interest in the data.

Kailkhura collaborates with Bremer and Thiagarajan on a project, funded by the Laboratory Directed Research and Development Program, aimed at exploring spectral sampling of HD spaces. In this context, *spectral* refers to the frequency of change among data points—a necessary consideration, the team argues, in addition to the data's spatial arrangement. Spectral analysis can enable better understanding of space-filling sample designs by finding a balance between uniform and random coverage. The project's goals are to determine optimal sampling patterns and to create ML algorithms that can generate those samples in any HD space.

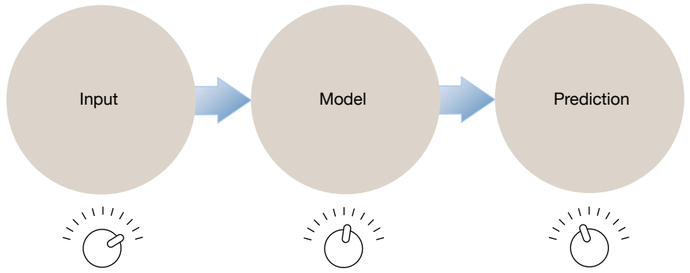
The project team uses a combination of exploration and exploitation techniques—sampling input variables independently of the output while using knowledge of the output to guide sample design, respectively. Project leader Thiagarajan notes, "We start the optimization process with blind exploration of data points, then switch to exploitation to search the regions of highest interest." This hybrid approach achieves better results than traditional methods by weighing both high- and low-frequency information—information respectively about more and less frequent change. The range of analyzable frequencies is maximized, providing statistically higher confidence in results. In recently published tests using data from NIF hot spot simulations, Livermore's spectral-sampling technique doubles the accuracy to significantly outperform other sample designs. Ultimately, optimized data inputs improve the ability of ML models to make useful predictions.

Essential Expertise

Examples abound of Lawrence Livermore's growing demand for machine learning (ML) to solve challenges in scientific data analysis. One research team created a toolkit that trains massive neural networks on image data. (See *S&TR*, June 2016, pp. 16–19.) Another project focuses on time-varying data, which reveal patterns in time. In this scenario, new ML algorithms progressively use existing observation-based data to forecast future events. For example, clinical decision making could be enhanced by ML analysis of trends in patient data.

One National Ignition Facility project leverages ML to analyze the largest-ever data set from inertial confinement fusion (ICF) implosions. (See *S&TR*, September 2018, pp. 16–19.) Another group is developing an innovative cognitive computing platform that combines ML with graph analytics and other areas of artificial intelligence to improve ICF simulation efficiency.

ML is also speeding up data analysis and prediction in three-dimensional printing and making multimodal data analysis easier in nuclear nonproliferation. Materials scientists use ML and big data analytics to accelerate materials synthesis and optimization. (See *S&TR*, July/August 2017, pp. 16–19.) ML technology helps Livermore scientists catalog and interpret objects orbiting Earth and process huge volumes of data captured by ground- and space-based telescopes. Livermore has also partnered with several institutions to accelerate drug discovery and development by integrating high-performance computing, ML, and other data science technologies. Computer scientist Rushil Anirudh notes, "The possibilities of ML are exciting. Whenever we reach a roadblock, we find ways to break through with ML."



Perturbation is one method of exploring an ML model's interpretability in which researchers adjust—in a manner akin to turning a dial—different factors and observe the effects.

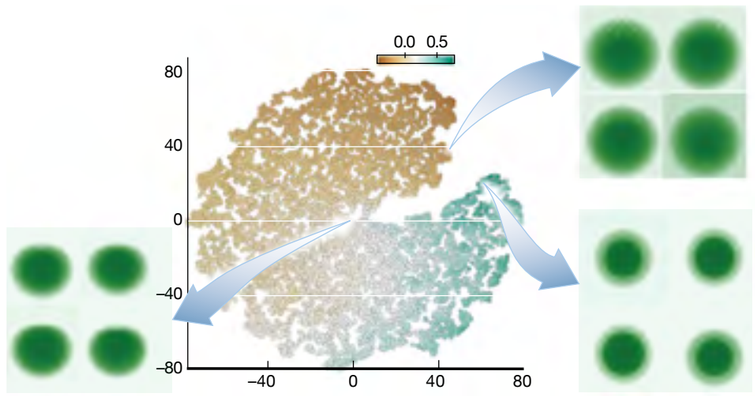
Trusting the Model

Model interpretability is another methodological ML pursuit at the Laboratory. "Human nature requires justification," states Liu. "We want to know which symptoms correlate to a particular diagnosis. We want to know how a conclusion was reached." Justification means providing rationale for how ML models work and the results they predict so stakeholders will trust both. Liu continues, "Interpretability is a necessary part of explaining or modifying an ML model, especially if the application is as important as NIF and not simply images of cats and dogs." Interpretability involves confronting tricky questions. For example, Bremer asks, "If a model is trained on a certain data set, how does one know it is not biased toward those data's properties?" An ML model might advise a bank that residents of a certain neighborhood are unsuitable candidates for a mortgage loan. If an applicant's address is the only criterion affecting loan approvals, then the model ignores other relevant information such as credit score or loan repayment history. Avoiding bias means understanding how the model arrives at a prediction and finding where bias might originate.

ML models do not have to specify a path to a solution. Consequently, Bremer cautions, "We may not understand how the model performed its analysis, which undermines confidence in the solution, especially for nonexperts." The stakes are even higher for large-scale models with thousands or millions of parameters. The quest for useful ML interpretation comes with many challenges, such as the absence of a universally agreed-upon explanation. To control error and variability in new ML approaches, Liu advocates for transparency so that the model is not merely a "black box."

Liu studies ML models through exploratory analysis. He states, "Conventional interpretation techniques study the model as an invariant object, where its behaviors are recorded and analyzed in an offline fashion." Instead, Liu recommends perturbation as one interpretation tool. Analogous to adjusting a radio's volume by turning a knob, researchers can perturb different variables and observe the behavior of others. For instance, masking a localized part of an image can affect the prediction of what the image contains. By dividing an image into a pixel grid and shifting the mask around the grid, researchers can calculate each pixel's importance in identifying the desired object. Liu explains, "This approach investigates relationships between inputs and outputs to determine which properties of the input contribute to the prediction."

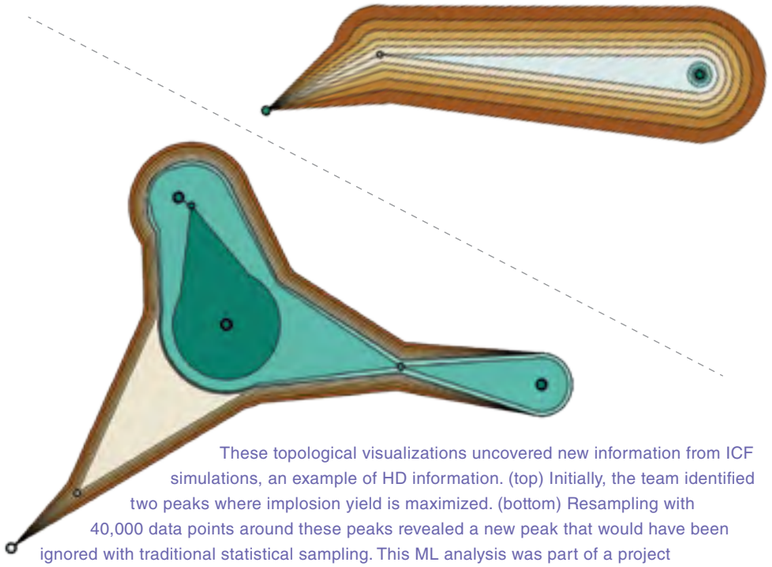
Optimization of latent spaces presents another step toward interpretability. Latent space lies between ML processing's encoding and decoding stages and captures variations and other key underlying information in a compressed representation of the data. Unsupervised ML models map inputs through layers of NNs into the latent space, where data are reduced into lower dimensional representations, enabling the model to identify hidden features beyond those observed. "In many real-world scenarios, HD data can be compressed into spaces with as few as two to four dimensions," notes Kailkhura, whose work in sample design optimizes an understanding of these latent spaces. Knowing more about the features of these latent spaces makes results more interpretable.



A 10-dimensional latent space of inertial confinement fusion (ICF) simulation data is reduced to the 2-dimensional visualization shown, in which the axes and scale no longer have explicit physical meaning. (insets) Different areas of the latent space capture various shapes of ICF images, providing insight into how the ML model interprets variations in high-dimensional (HD) data.

"Latent spaces are compact and descriptive but typically not transparent or intuitive," says Bremer. Therefore, Liu and colleagues apply nonlinear dimensionality reduction functions to latent spaces and use visualizations to discover feature variations captured and distributed throughout these spaces. By comparing visual encodings of the HD space, researchers can determine how many dimensions yield the most valuable information. In one study of ICF simulations, the team compared image patterns in 10- and 16-dimensional latent spaces and found that the latter did not fully use all dimensions. Liu summarizes, "By reducing the dimensions when exploring the latent space, we can directly assess the information captured by that space and explain the differences between simulations."

For many scientists who rely on ML, seeing is believing. Topological data analysis is another valuable tool for understanding the structure of HD spaces, and the resulting visualizations help Livermore researchers explain and verify ML models. "Topology produces abstract structures that generalize to high dimensions," notes Bremer. Laboratory researchers have released open-source software that render data relationships through mountains, valleys, and other maplike contours. Bremer continues, "We can extract HD properties and show them as a low-dimensional terrain.



These topological visualizations uncovered new information from ICF simulations, an example of HD information. (top) Initially, the team identified two peaks where implosion yield is maximized. (bottom) Resampling with 40,000 data points around these peaks revealed a new peak that would have been ignored with traditional statistical sampling. This ML analysis was part of a project investigating ICF target shape. (See *S&TR*, September 2018, pp. 16–19.)

Visualizations allow us to find patterns or anomalies that other statistical methods may not find, so we can evaluate information that would otherwise be incomprehensible."

Case Study: Multiscale Modeling

In 2016, the Department of Energy (DOE) and the National Cancer Institute launched a multiyear partnership to advance cancer research using modern HPC resources. Livermore plays a central role in the program's three pilot projects. (See *S&TR*, October/November 2016, pp. 4–11.) One project, nicknamed Pilot 2, brings together three DOE laboratories—Lawrence Livermore, Los Alamos, and Oak Ridge—and Frederick National Laboratory for Cancer Research to explain interactions between cell membranes and specific proteins that induce many forms of cancer. (Pilots 1 and 3 focus on drug discovery and patient health records.)

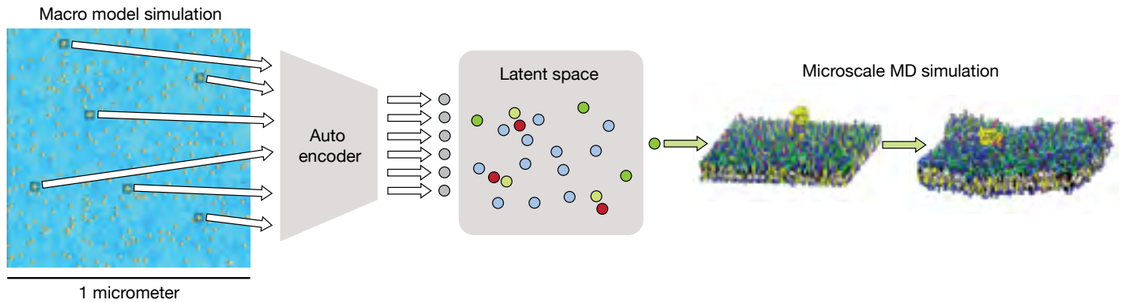
To guide multiscale simulations of these interactions, Pilot 2 collaborators—including Bhatia and Bremer—develop ML approaches aiming to understand both the mechanism of a protein called RAS and the signaling chain that causes another protein, RAF, to interact with RAS. "ML is at the very center of this project, integrating different areas of expertise," states Bhatia. "We use ML to locate phenomena occurring on the cell membrane in coarse simulations, which we can then investigate more closely with higher fidelity simulations." With this computational steering approach, researchers guide simulations to gain specific insight while maximizing the throughput of computational resources.

Understanding protein biology requires modeling at different spatial and temporal scales—from nano- to milliseconds and from nano- to micrometers. Bhatia explains, "Simulating the underlying phenomena with sufficient accuracy at fine scales is prohibitively expensive computationally." Therefore, the project's sophisticated ML model is trained on coarse macroscale simulations before resources are spent on more detailed microscale molecular dynamics (MD) simulations. He continues, "Coarse simulations give us a reasonable approximation of results. The ML model identifies important areas, such as a small location where a protein

interacts with the cell membrane, where we should invest our resources at a higher resolution. We want to investigate enough regions of interest to make statistical claims over a long temporal range without running simulations for the entire period." This tactic could cut simulation time from months to days.

Computational steering is a sampling problem. Accordingly, the approach takes advantage of latent spaces. Bhatia says, "We have millions of potentially interesting data points with nonlinear, highly complex relationships. For example, consider finding similar-looking houses among millions of photos. We could not simply compare individual pixels to determine similarity." The team's ML solution includes an autoencoder—a deep NN—that reduces the data into a latent space. From there, the model chooses the features most dissimilar from previous iterations and ranks the results according to importance—from most to least anomalous. Even with compression, a million data points could be flagged, which is why using latent spaces is key.

In a process called adaptive sampling, data generated by macroscale simulations inform sampling of the MD simulations—the latter, in turn, becomes part of the feedback loop to update the former. Together, autoencoding, adaptive sampling, and the in situ feedback cycle allow the team to manage over a million samples through HD analysis and, therefore, run macro simulations with the accuracy of MD. "These types of simulations are novel, and we are scaling the workflow to target a supercomputer such as Sierra," states Bhatia. In 2018, the Pilot 2 team reached a major milestone by computationally steering such multiscale simulations on Sierra.



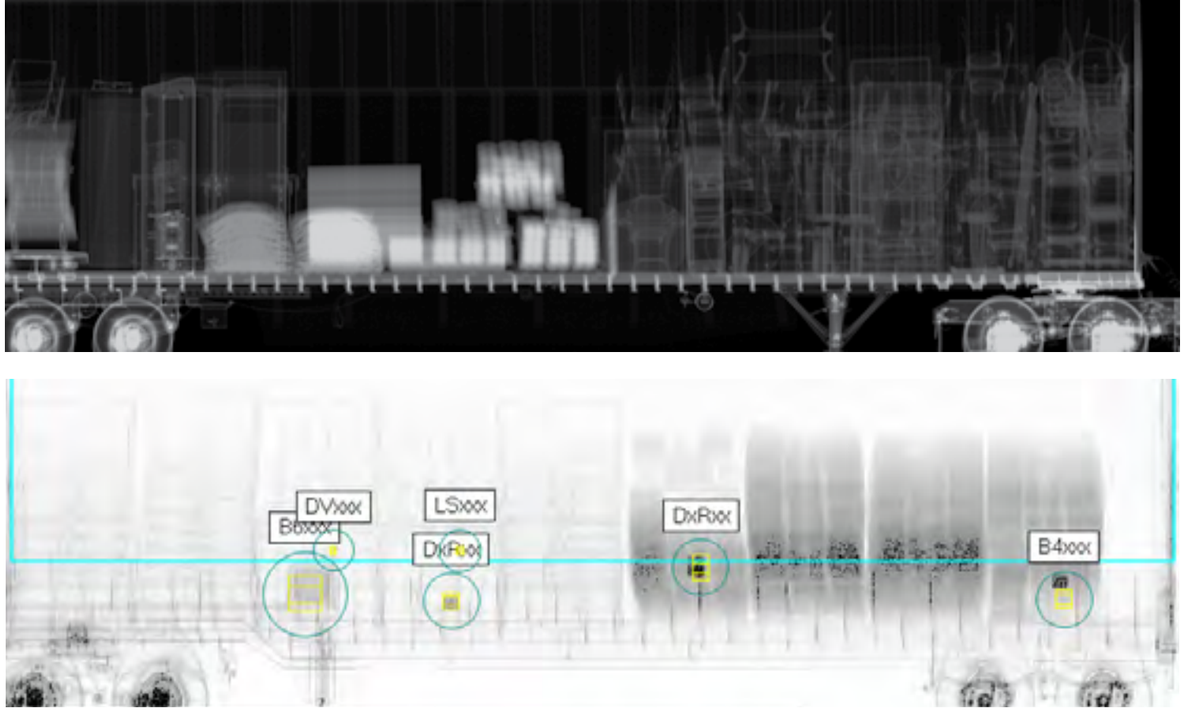
An illustration shows how ML is at the center of the Pilot 2 cancer research workflow, connecting coarse macroscale simulations (left) and fine microscale molecular dynamics (MD) simulations (right) used to investigate mediation of cancer initiation by the protein RAS. An autoencoder reduces macroscale model data into a latent space, where the data are ranked by novelty and importance. MD simulations are reserved for the most important regions, thereby conserving computational resources. The result is simulations with macro (long) length and timescales that also provide insights at the microscale.

Case Study: Threat Detection

Three screening scenarios—medical diagnosis, airport luggage, and commercial truck cargo—share important characteristics. All require expert analysis of image scans, and threats are reduced with quick identification of suspicious objects. Automatically flagging suspicious areas in an image saves human operators' time while minimizing errors. For example, maximized information from a computed tomography scan can help reduce a patient's radiation exposure or improve prognosis with early cancer detection. In all three scenarios, the goal is higher detection rates with fewer false alarms.

Lung cancer nodules are inconsistent in size and shape and may not appear clearly in a lung image. A radiologist can mark a nodule in an image but cannot be expected to label every affected pixel. ML algorithms require more specific coordinates for nodule location, so the model must learn to create detailed labels at different stages of analysis. Anirudh explains, "We use unsupervised strategies to estimate nodule characteristics, such as boundaries, in weakly labeled data." Kim adds, "This model will not replace radiologists' expertise but will significantly reduce their workloads by filtering out images that do not need close review. The model can also provide a 'second opinion' to reduce diagnostic errors."

Luggage screening technology stands to benefit from ML-driven efficiencies in image quality, as airport scanners often provide sparse views of imaged objects. The Livermore team built a system of one- and two-dimensional NNs to recover limited-angle or partial-view images. Mindful of interpretability, they also designed a confidence score to gauge results reliability. The score is calculated from estimates of pixel variabilities within the model's latent space and is correlated with reconstruction quality. The team's image reconstruction and segmentation techniques have shown higher fidelity to ground truth than other methods.



(top) Illicit materials inside a vehicle are difficult to discern from this single-view scan. (bottom) The Laboratory's ML source-separation technique divides the image into layers from which the three-dimensional depth of the contents can be discovered.

At ports of entry, analysts see only a two-dimensional scan and must decide whether cargo contains, for example, nuclear materials stashed among a truckload of appliances. The cargo's three-dimensional depth cannot be directly observed and so is inferred from a sum of the layers in a single image. The Livermore team has developed a source-separation model that splits a single image into multiple images to predict distribution of cargo materials. By training on probabilistic "clean" data separated into layers, this unsupervised ML model develops a surrogate for physical materials, then applies it to subsequent scans. Thiagarajan compares this technique to the way the brain identifies merged objects, saying, "If I show you separate images of a face and a pair of sunglasses, you can mentally combine them."

In addition, Livermore researchers are moving toward what Anirudh calls "ML 2.0"—a more robust unsupervised model that does not collect data sets for every task. For example, Kim explains, "Thousands of unlabeled bags are scanned daily at airports. When a new object is introduced, the scanner needs to detect the abnormality for the security officer to investigate." The team is solving such inverse problems with adversarial NNs, which perform generative and discriminative evaluations to reduce errors.

Disruptive Advances

In the quest to understand human intelligence, researchers across the Laboratory are evolving scientific ML in areas such as mathematical neuroscience, brain-inspired network architectures, representation learning, and multistage training algorithms. Thiagarajan says, "Combining scientific exploration and artificial intelligence opens up exciting opportunities for solving real-world challenges."

Livermore's ML experts agree that most research teams at the Laboratory will eventually seek ML-driven solutions to the challenges they face. In fact, many mission-critical programs already rely on ML technologies. Kailkhura states, "Once you grasp the concepts, the applications are numerous." Bremer adds, "Grand challenges in science and computing cannot be addressed with incremental improvements. Instead, we must look for disruptive advances with significant technical, programmatic, and strategic impact. Livermore is absolutely the right place—perhaps one of the only places—to do this."

—Holly Auten

Key Words: algorithm, computational steering, deep learning, high-dimensional (HD) space, high-performance computing (HPC), image reconstruction, image segmentation, inertial confinement fusion (ICF), Laboratory Directed Research and Development Program, latent space, machine learning (ML), model interpretability, neural network (NN), sample design, simulation, source separation, spectral sampling, topological visualization.

For further information contact Peer-Timo Bremer (925) 422-7365 (bremer5@llnl.gov).