REVIEW

Machine learning for data-driven discovery in solid Earth geoscience

KARIANNE J. BERGEN, PAUL A. JOHNSON, [...], AND GREGORY C. BEROZA +1 authors Authors Info & Affiliations

SCIENCE 22 Mar 2019 Vol 363, Issue 6433 DOI: 10.1126/science.aau0323

Automating geoscience analysis

Solid Earth geoscience is a field that has very large set of observations, which are ideal for analysis with machine-learning methods. Bergen *et al.* review how these methods can be applied to solid Earth datasets. Adopting machine-learning techniques is important for extracting information and for understanding the increasing amount of complex data collected in the geosciences.

Science, this issue p. eaau0323

Structured Abstract

BACKGROUND

The solid Earth, oceans, and atmosphere together form a complex interacting geosystem. Processes relevant to understanding Earth's geosystem behavior range in spatial scale from the atomic to the planetary, and in temporal scale from milliseconds to billions of years. Physical, chemical, and biological processes interact and have substantial influence on this complex geosystem, and humans interact with it in ways that are increasingly consequential to the future of both the natural world and civilization as the finiteness of Earth becomes increasingly apparent and limits on available energy, mineral resources, and fresh water increasingly affect the human condition. Earth is subject to a variety of geohazards that are poorly understood, yet increasingly impactful as our exposure grows through increasing urbanization, particularly in hazard-prone areas. We have a fundamental need to develop the best possible predictive understanding of how the geosystem works, and that understanding must be informed by both the present and the deep past. This understanding will come through the analysis of increasingly large geo-datasets and from computationally intensive simulations, often connected through inverse problems. Geoscientists are faced with the challenge of extracting as much useful information as possible and gaining new insights from these data, simulations, and the interplay between the two. Techniques from the rapidly evolving field of machine learning (ML) will play a key role in this effort.

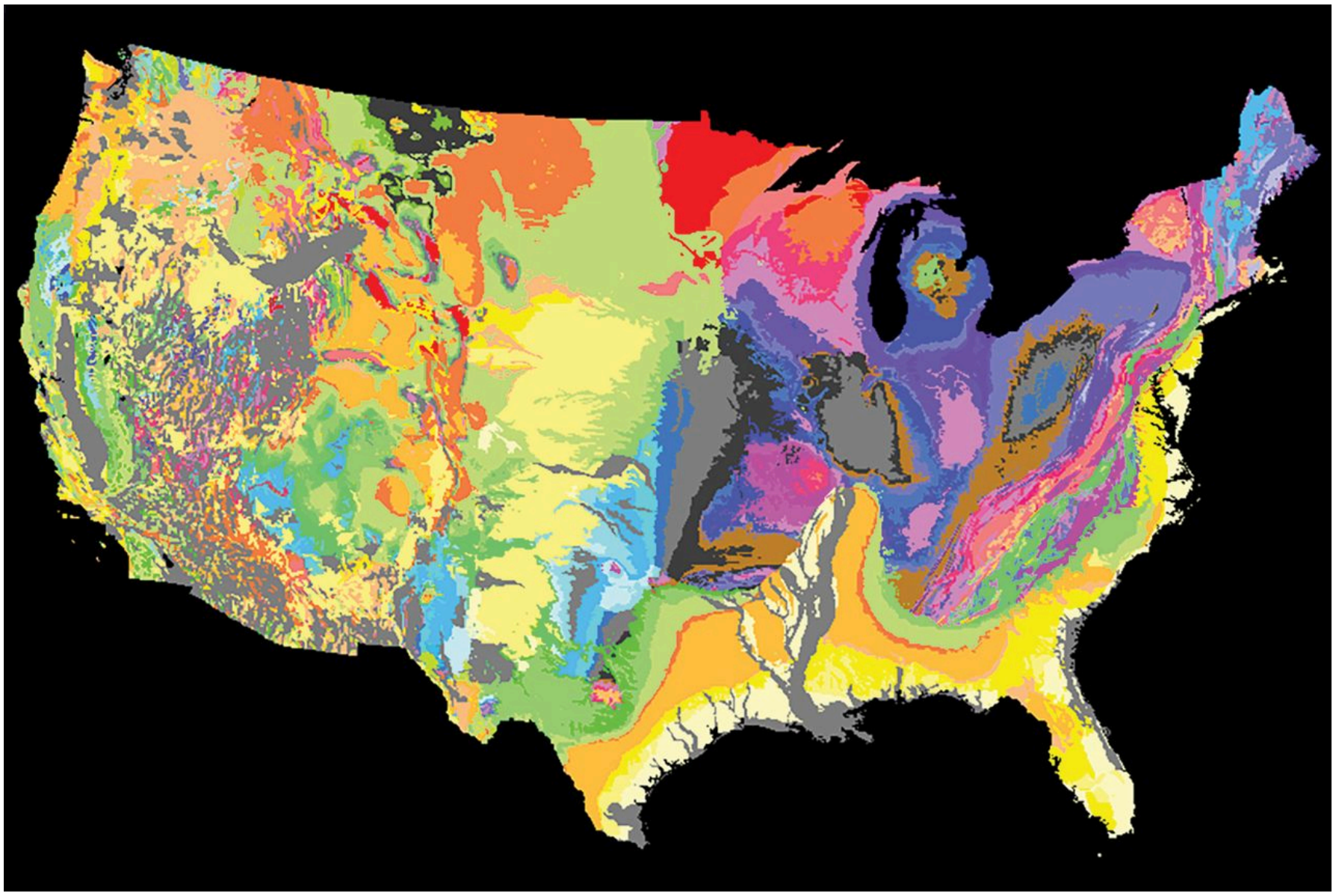
ADVANCES

The confluence of ultrafast computers with large memory, rapid progress in ML algorithms, and the ready availability of large datasets place geoscience at the threshold of dramatic progress. We anticipate that this progress will come from the application of ML across three categories of research effort: (i) automation to perform a complex prediction task that cannot easily be described by a set of explicit commands; (ii) modeling and inverse problems to create a representation that approximates numerical simulations or captures relationships; and (iii) discovery to reveal new and often unanticipated patterns, structures, or relationships. Examples of automation include geologic mapping using remote-sensing data, characterizing the topology of fracture systems to model subsurface transport, and classifying volcanic ash particles to infer eruptive mechanism. Examples of modeling include approximating the viscoelastic response for complex rheology, determining wave speed models directly from tomographic data, and classifying diverse seismic events. Examples of discovery include predicting laboratory slip events using observations of acoustic emissions, detecting weak earthquake signals using similarity search, and determining the connectivity of subsurface reservoirs using groundwater tracer observations.

OUTLOOK

The use of ML in solid Earth geosciences is growing rapidly, but is still in its early stages and making uneven progress. Much remains to be done with existing datasets from long-standing data sources, which in many cases are largely unexplored. Newer, unconventional data sources such as light detection and ranging (LiDAR), fiber-optic sensing, and crowd-sourced measurements may demand new approaches through both the volume and the character of information that they present.

Practical steps could accelerate and broaden the use of ML in the geosciences. Wider adoption of open-science principles such as open source code, open data, and open access will better position the solid Earth community to take advantage of rapid developments in ML and artificial intelligence. Benchmark datasets and challenge problems have played an important role in driving progress in artificial intelligence research by enabling rigorous performance comparison and could play a similar role in the geosciences. Testing on high-quality datasets produces better models, and benchmark datasets make these data widely available to the research community. They also help recruit expertise from allied disciplines. Close collaboration between geoscientists and ML researchers will aid in making quick progress in ML geoscience applications. Extracting maximum value from geoscientific data will require new approaches for combining data-driven methods, physical modeling, and algorithms capable of learning with limited, weak, or biased labels. Funding opportunities that target the intersection of these disciplines, as well as a greater component of data science and ML education in the geosciences, could help bring this effort to fruition.



Digital geology.
Digital representation of the geology of the conterminous United States. [Geology of the Conterminous United States at 1:2,500,000 scale; a digital representation of the 1974 P. B. King and H. M. Beikman map by P. G. Schruben, R. E. Arndt, W. J. Bawiec]

The list of author affiliations is available in the full article online.

Abstract

Understanding the behavior of Earth through the diverse fields of the solid Earth geosciences is an increasingly important task. It is made challenging by the complex, interacting, and multiscale processes needed to understand Earth's behavior and by the inaccessibility of nearly all of Earth's subsurface to direct observation. Substantial increases in data availability and in the increasingly realistic character of computer simulations hold promise for accelerating progress, but developing a deeper understanding based on these capabilities is itself challenging. Machine learning will play a key role in this effort. We review the state of the field and make recommendations for how progress might be broadened and accelerated.

The solid Earth, oceans, and atmosphere together form a complex interacting geosystem. Processes relevant to understanding its behavior range in spatial scale from the atomic to the planetary, and in temporal scale from milliseconds to billions of years. Physical, chemical, and biological processes interact and have substantial influence on this complex geosystem. Humans interact with it too, in ways that are increasingly consequential to the future of both the natural world and civilization as the finiteness of Earth becomes increasingly apparent and limits on available energy, mineral resources, and fresh water increasingly affect the human condition. Earth is subject to a variety of geohazards that are poorly understood, yet increasingly impactful as our exposure grows through increasing urbanization, particularly in hazard-prone areas. We have a fundamental need to develop the best possi-

ble predictive understanding of how the geosystem works, and that understanding must be informed by both the present and the deep past.

In this review we focus on the solid Earth. Understanding the material properties, chemistry, mineral physics, and dynamics of the solid Earth is a fascinating subject, and essential to meeting the challenges of energy, water, and resilience to natural hazards that humanity faces in the 21st century. Efforts to understand the solid Earth are challenged by the fact that nearly all of Earth's interior is, and will remain, inaccessible to direct observation. Knowledge of interior properties and processes are based on measurements taken at or near the surface, are discrete, and are limited by natural obstructions such that aspects of that knowledge are not constrained by direct measurement.

For this reason, solid Earth geoscience (sEg) is both a data-driven and a model-driven field with inverse problems often connecting the two. Unanticipated discoveries increasingly will come from the analysis of large datasets, new developments in inverse theory, and procedures enabled by computationally intensive simulations. Over the past decade, the amount of data available to geoscientists has grown enormously, through larger deployments of traditional sensors and through new data sources and sensing modes. Computer simulations of Earth processes are rapidly increasing in scale and sophistication such that they are increasingly realistic and relevant to predicting Earth's behavior. Among the foremost challenges facing geoscientists is how to extract as much useful information as possible and how to gain new insights from both data and simulations and the interplay between the two. We argue that machine learning (ML) will play a key role in that effort.

ML-driven breakthroughs have come initially in traditional fields such as computer vision and natural language processing, but scientists in other domains have rapidly adopted and extended these techniques to enable discovery more broadly (*1*–*4*). The recent interest in ML among geoscientists initially focused on automated analysis of large datasets, but has expanded into the use of ML to reach a deeper understanding of coupled processes through data-driven discoveries and model-driven insights. In this review we introduce the challenges faced by the geosciences, present emerging trends in geoscience research, and provide recommendations to help accelerate progress.

ML offers a set of tools to extract knowledge and draw inferences from data (*5*). It can also be thought of as the means to artificial intelligence (AI) (*6*), which involves machines that can perform tasks characteristic of human intelligence (*7*, *8*). ML algorithms are designed to learn from experience and recognize complex patterns and relationships in data. ML methods take a different approach to analyzing data than classical analysis techniques (Fig. 1)—an approach that is robust, fast, and allows exploration of a large function space (Fig. 2).

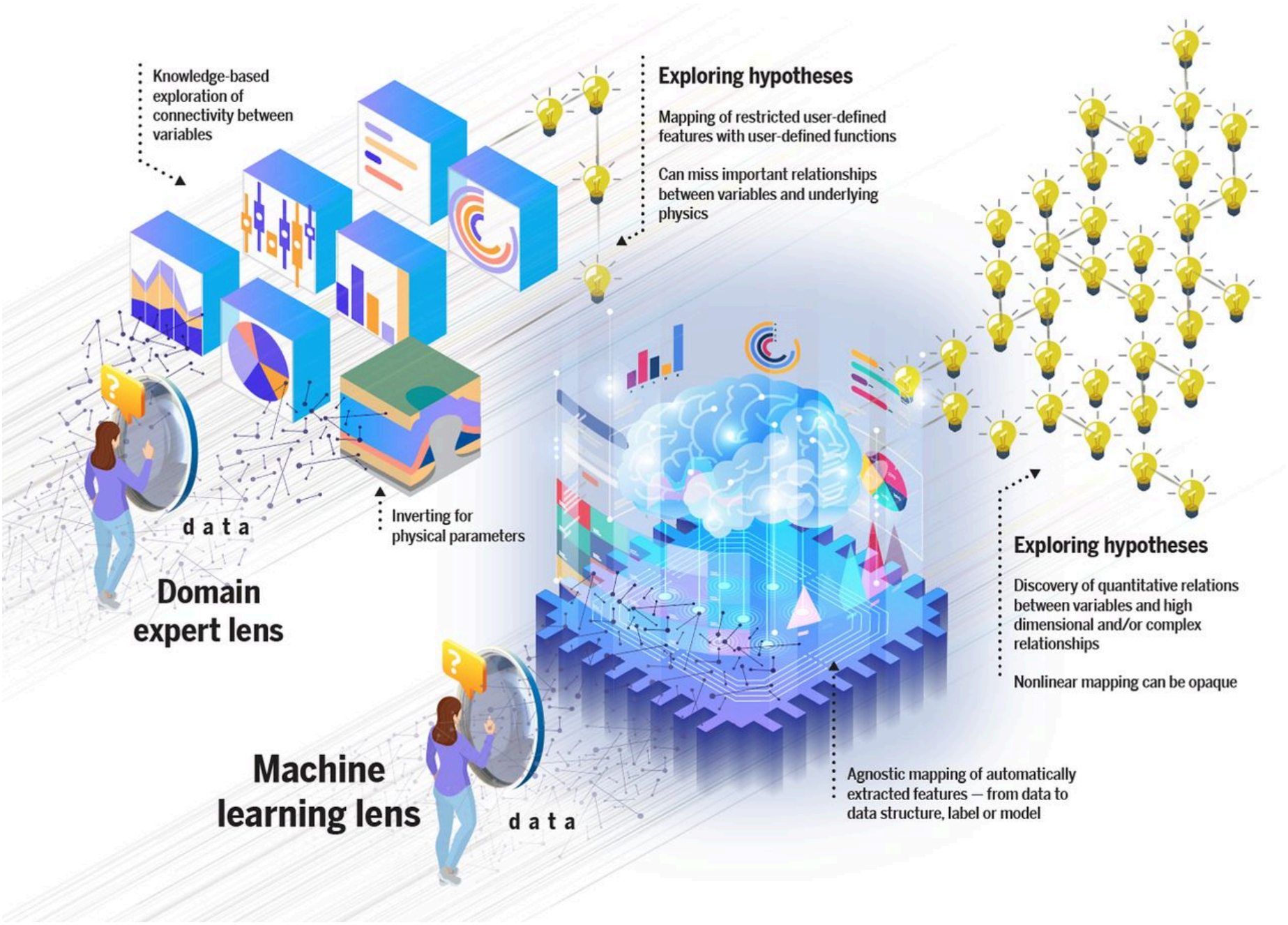


Fig. 1 How scientists analyze data: the conventional versus the ML lens for scientific analysis.
ML is akin to looking at the data through a new lens. Conventional approaches applied by domain experts (e.g., Fourier analysis) are preselected and test a hypothesis or simply display data differently. ML explores a larger function space that can connect data to some target or label. In doing so, it provides the means to discover relations between variables in high-dimensional space. Whereas some ML approaches are transparent in how they find the function and mapping, others are opaque. Matching an appropriate ML approach to the problem is therefore extremely important.

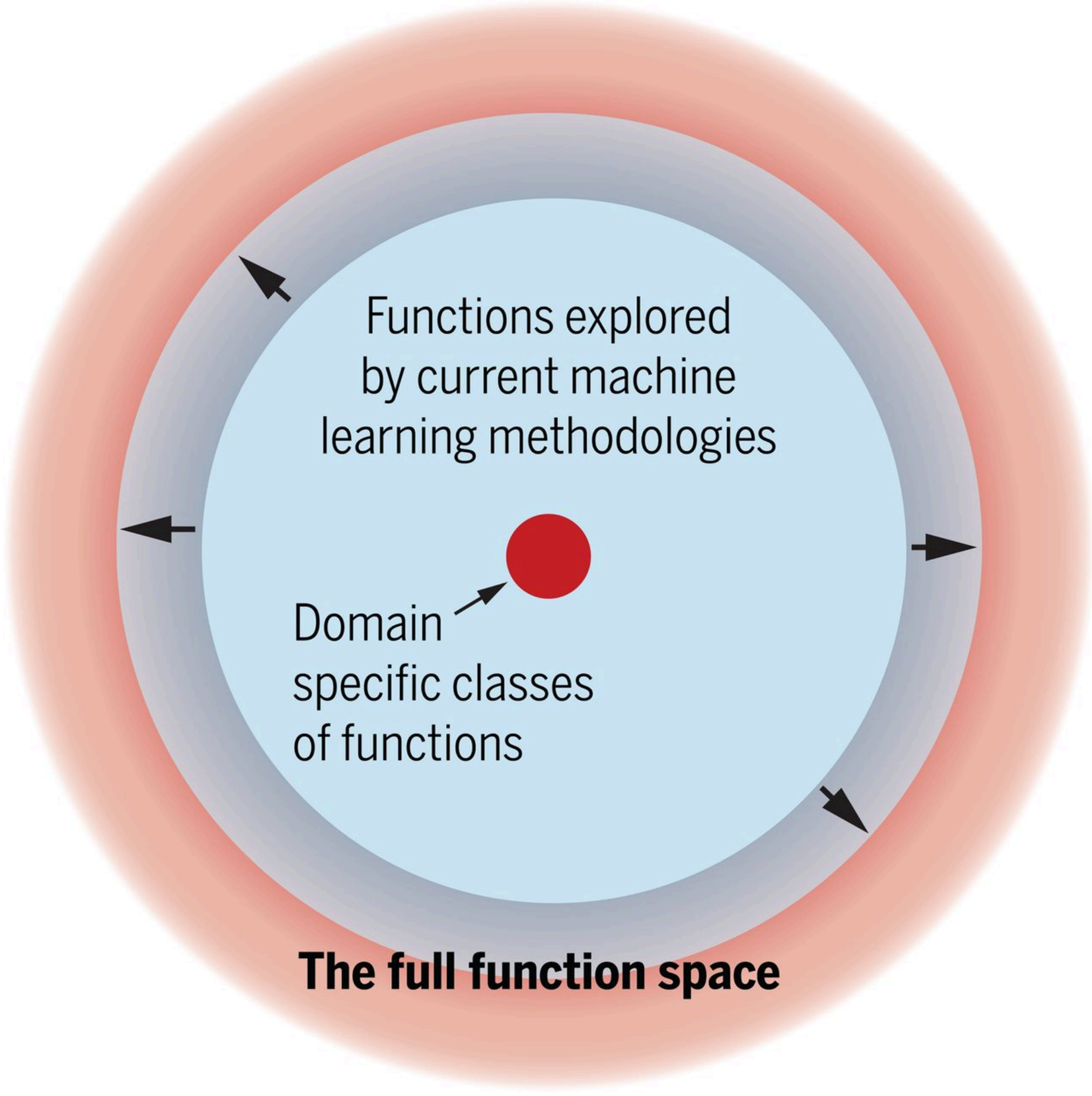


Fig. 2 The function space used by domain experts and that used by ML.
The function space of user-defined functions employed by scientists, in contrast to the functional space used by ML, is contained within the entire possible function space. The function space that ML employs is expanding rapidly as the computational costs and runtimes decrease and memory, depths of networks, and available data increase.

The two primary classes of ML algorithms are supervised and unsupervised techniques. In supervised learning, the ML algorithm "learns" to recognize a pattern or make general predictions using known examples. Supervised learning algorithms create a map, or model, f that relates a data (or feature) vector x to a corresponding label or target vector y: $y = f(x)$, using labeled training data [data for which both the input and corresponding label $(x^{(i)}, y^{(i)})$ are known and available to the algorithm] to optimize the model. For example, a supervised ML classifier might learn to detect cancer in medical images using a set of physician-annotated examples (*9*). A well-trained model should be able to generalize and make accurate predictions for previously unseen inputs (e.g., label medical images from new patients).

Unsupervised learning methods learn patterns or structure in datasets without relying on label characteristics. In a well-known example, researchers at Google's X lab developed a feature-detection algorithm that learned to recognize cats after being exposed to millions of images from YouTube without prompting or prior information about cats (*10*). Unsupervised learning is often used for exploratory data analysis or visualization in datasets for which no or few labels are available, and includes dimensionality reduction and clustering.

The many different algorithms for supervised and unsupervised learning each have relative strengths and weaknesses. The algorithm choice depends on a number of factors including (i) availability of labeled data, (ii) dimensionality of the data vector, (iii) size of dataset, (iv) continuous- versus discrete-valued prediction target, and (v) desired model interpretability. The level of model interpretability may be of particular concern in geoscientific applications. Although interpretability may not be necessary in a highly accurate image recognition system, it is critical when the goal is to gain physical insight into the system.

Machine learning in solid Earth geosciences

Scientists have been applying ML techniques to problems in the sEg for decades (*11*–*13*). Despite the promise shown by early proof-of-concept studies, the community has been slow to adopt ML more broadly. This is changing rapidly. Recent performance breakthroughs in ML, including advances in deep learning and the availability of powerful, easy-to-use ML toolboxes, have led to renewed interest in ML among geoscientists. In sEg, researchers have leveraged ML to tackle a diverse range of tasks that we group into the three interconnected modes of automation, modeling and inverse problems, and discovery (Fig. 3).

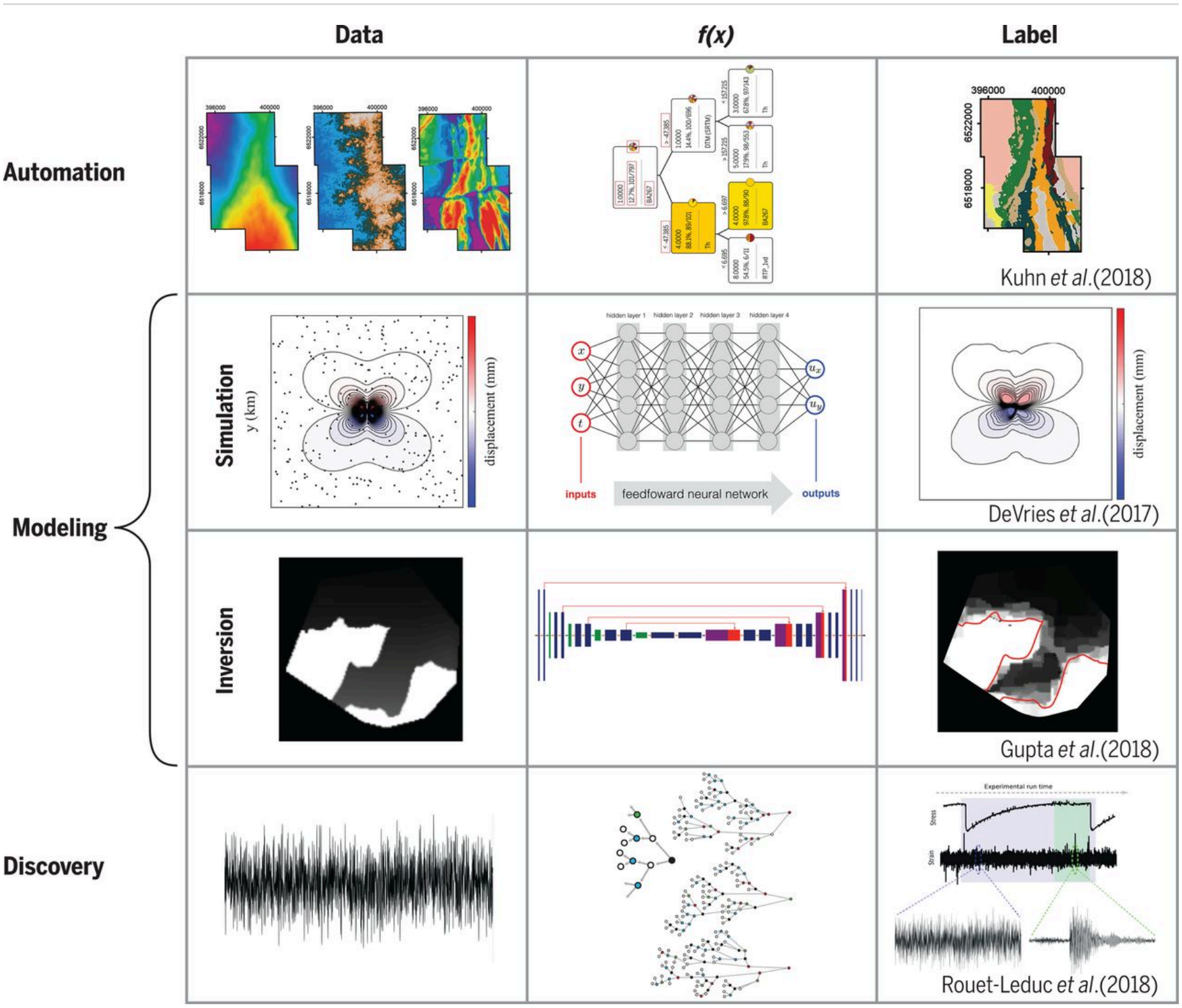


Fig. 3 Common modes of ML. The top row shows an example of an automated approach to mapping lithology using remote-sensing data by applying a random forest ML approach. This approach works well with sparse ground-truth data and gives robust estimates of the uncertainty of the predicted lithology (*35*). The second row shows training a deep neural network to learn a computationally efficient representation of viscoelastic solutions in Earth, allowing calculations to be done quickly, reliably, and with high spatial and temporal resolutions (*23*). The third row shows an example of inversion where the input is a nonnegative least-squares reconstruction and the network is trained to reconstruct a

Automation is the use of ML to perform a complex task that cannot easily be described by a set of explicit commands. In automation tasks, ML is selected primarily as a tool for making highly accurate predictions (or labeling data), particularly when the task is difficult for humans to perform or explain. Examples of ML used for automation outside the geosciences include image recognition (*14*) or movie recommendation (*15*) systems. ML can improve upon expert-designed algorithms by automatically identifying better solutions among a larger set of possibilities. Automation takes advantage of a strength of ML algorithms—their ability to process and extract patterns from large or high-dimensional datasets—to replicate or exceed human performance. In the sEg, ML is used to automate the steps in large-scale data analysis pipelines, as in earthquake detection (*16*) or earthquake early warning (*17*–*19*), and to perform specialized, repetitive tasks that would otherwise require time-consuming expert analysts, such as categorizing volcanic ash particles (*20*).

ML can also be used for modeling, or creating a representation that captures relationships and structure in a dataset. This can take the form of building a model to represent complex, unknown, or incompletely understood relationships between data and target variables; e.g., the relationship between earthquake source parameters and peak ground acceleration for ground motion prediction (*21*, *22*). ML can also be used to build approximate or surrogate models to speed large computations, including numerical simulations (*23*, *24*) and inversion (*25*). Inverse problems connect observational data, computational models, and physics to enable inference about physical systems in the geosciences. ML, especially deep learning, can aid in the analysis of inverse problems (*26*). Deep neural networks, with architectures informed by the inverse problem itself, can learn an inverse map for critical speedups over traditional reconstructions, and the analysis of generalization of ML models can provide insights into the ill-posedness of an inverse problem.

Data-driven discovery, the ability to extract new information from data, is one of most exciting capabilities of ML for scientific applications. ML provides scientists with a set of tools for discovering new patterns, structure, and relationships in scientific datasets that are not easily revealed through conventional techniques. ML can reveal previously unidentified signals or physical processes (*27*–*31*), and extract key features for representing, interpreting, or visualizing data (*32*–*34*). ML can help to minimize bias—for example, by discovering patterns that are counter-intuitive or unexpected (*29*). It can also be used to guide the design of experiments or future data collection (*35*).

These themes are all interrelated; modeling and inversion can also provide the capability for automated predictions, and the use of ML for automation, modeling, or inversion may yield new insights and fundamental discoveries.

Methods and trends for supervised learning

Supervised learning methods use a collection of examples (training data) to learn relationships and build models that are predictive for previously unseen data. Supervised learning is a powerful set of tools that have successfully been used in applications spanning the themes of automation, modeling and inversion, and discovery (Fig. 4). In this section we organize recent supervised learning applications in the sEg by ML algorithm, which we order roughly by model complexity, starting with the relatively simple logistic regression classifier and ending with deep neural networks. In general, more complex models require more training data and less feature engineering.

Logistic regression

Logistic regression (*36*) is a simple binary classifier that estimates the probability that a new data point belongs to one of two classes. Reynen and Audet (*37*) apply a logistic regression classifier to distinguish automatically between earthquake signals and explosions, using polarization and frequency features extracted from seismic waveform data. They extend their approach to detect earthquakes in continuous data by classifying each time segment as earthquake or noise. They used class probabilities at each seismic station to combine the detection results from multiple stations in the seismic network. Pawley *et al*. (*38*) use logistic regression to separate aseismic from seismogenic injection wells for induced seismicity, using features in the model to identify geologic factors, including proximity of the well to basement, associated with a higher risk of induced seismicity.

Graphical models

Many datasets in the geosciences have a temporal component, such as the ground motion time-series data recorded by seismometers. Although most ML algorithms can be adapted for use on temporal data, some methods, like graphical models, can directly model temporal dependencies. For example, hidden Markov models (HMMs) are a technique for modeling sequential data and have been widely used in speech recognition (*39*). HMMs have been applied to continuous seismic data for the detection and classification of alpine rockslides (*40*), volcanic signals (*41*, *42*), regional earthquakes (*43*), and induced earthquakes (*44*). A detailed explanation of HMMs and their application to seismic waveform data can be found in Hammer *et al*. (*42*). Dynamic Bayesian networks (DBNs), another type of graphical model that generalizes HMMs, have also been used for earthquake detection (*45*, *46*). In exploration geophysics, hierarchical graphical models have been applied to determine the connectivity of subsurface reservoirs from time-series measurements using priors derived from convection-diffusion equations (*47*). The authors report that use of a physics-based prior is key to obtaining a reliable model. Graph-based ML emulators were used by Srinivasan *et al*. (*48*) to mimic high-performance physics-based computations of flow through fracture networks, making robust uncertainty quantification of fractured systems possible (*48*).

Support vector machine

Support vector machine (SVM) is a binary classification algorithm that identifies the optimal boundary between the training data from two classes (*49*). SVMs use kernel functions, similarity functions that generalize the inner product, to enable an implicit mapping of the data into a higher-dimensional feature space. SVMs with linear kernels separate classes with a hyperplane, whereas nonlinear kernel functions allow for nonlinear decision boundaries between classes [see Cracknell and Reading (*50*) and Shahnas *et al*. (*51*) for explanations of SVMs and kernel methods, respectively].

Shahnas *et al*. (*51*) use an SVM to study mantle convection processes by solving the inverse problem of estimating mantle density anomalies from the temperature field. Temperature fields computed by numerical simulations of mantle convection are used as training data. The authors also train an SVM to predict the degree of mantle flow stagnation. Both support vector machines (*18*) and support vector regression (*19*) have been used for rapid magnitude estimation of seismic events for earthquake early warning. Support vector machines have also been used for discrimination of earthquakes and explosions (*52*) and for earthquake detection in continuous seismic data (*53*).

Random forests and ensemble learning

Decision trees are a supervised method for classification and regression that learn a piecewise-constant function, equivalent to a series of if-then rules that can be visualized by a binary tree structure. A random forest (RF) is an ensemble learning algorithm that can learn complex relationships by voting among a collection (“forest”) of randomized decision trees (*54*) [see Cracknell and Reading (*50*) for a detailed description of RFs]. Random forests are relatively easy to use and interpret. These are important advantages over methods that are opaque or require tuning many hyperparameters (e.g., neural networks, described below), and have contributed to the broad application of RFs within sEg.

Kuhn *et al*. (*35*) produced lithological maps in Western Australia using geophysical and remote-sensing data that were trained on a small subset of the ground area. Cracknell and Reading (*55*) found that random forests provided the best performance for geological mapping by comparing multiple supervised ML algorithms. Random forest predictions also improved three-dimensional (3D) geological models using remotely sensed geophysical data to constrain geophysical inversions (*56*).

Trugman and Shearer (*21*) discern a predictive relationship between stress drop and peak ground acceleration using RFs to learn nonlinear, nonparametric ground motion prediction equations (GMPEs) from a dataset of moderate-magnitude events in northern California. This departed from the typical use of linear regression to model the relationship between expected peak ground velocity or acceleration and earthquake site and source parameters that define GMPEs.

Valera *et al*. (*24*) used ML to characterize the topology of fracture patterns in the subsurface for modeling flow and transport. A graph representation of discrete fracture networks allowed RF and SVMs to identify subnetworks that

characterize the network flow and transport of the full network. The reduced network representations greatly decreased the computational effort required to estimate system behavior.

Rouet-Leduc *et al*. (*28*, *29*) trained a RF on continuous acoustic emission in a laboratory shear experiment to determine instantaneous friction and to predict time-to-failure. From the continuous acoustic data using the same laboratory apparatus, Hulbert *et al*. (*30*) apply a decision tree approach to determine the instantaneous fault friction and displacement on the laboratory fault. Rouet-Leduc *et al*. (*31*) scaled the approach to Cascadia by applying instantaneous seismic data to predict the instantaneous displacement rate on the subducting plate interface using GPS data as the label. In the laboratory and the field study in Cascadia, ML revealed unknown signals. Of interest is that the same features apply both at laboratory and field scale to infer fault physics, suggesting a universality across systems and scales.

Neural networks

Artificial neural networks (ANNs) are an algorithm loosely modeled on the interconnected networks of biological neurons in the brain (*57*). ANN models are represented as a set of nodes (neurons) connected by a set of weights. Each node takes a weighted linear combination of values from the previous layer and applies a nonlinear function to produce a single value that is passed to the next layer. "Shallow" networks contain an input layer (data), a single hidden layer, and an output layer (predicted response). Valentine and Woodhouse (*58*) present a detailed explanation of ANNs and the process of learning weights from training data. ANNs can be used for both regression and classification, depending on the choice of output layer.

ANNs have a long history of use in the geosciences [see (*59*, *60*) for reviews of early work], and they remain popular for modeling nonlinear relationships in a range of geoscience applications. De Wit *et al*. (*61*) estimate both the 1D P-wave velocity structure and model uncertainties from P-wave travel-time data by solving the Bayesian inverse problem with an ANN. This neural network–based approach is an alternative to using the standard Monte Carlo sampling approach for Bayesian inference. Käufl *et al*. (*25*) built an ANN model that estimates source parameters from strong motion data. The ANN model performs rapid inversion for source parameters in real time by precomputing computationally intensive simulations that are then used to train the neural network model.

ANNs have been used to estimate short-period response spectra (*62*), to model ground motion prediction equations (*22*), to assess data quality for focal mechanism and hypocenter location (*58*), and to perform noise tomography (*63*). Logistic regression and ANN models allowed Mousavi *et al*. (*64*) to characterize the source depth of microseimic events induced by underground collapse and sinkhole formation. Kong *et al*. (*17*) use an ANN with a small number of easy-to-compute features for use on a smartphone-based seismic network to distinguish between earthquake motion and motion due to user activity.

Deep neural networks

Deep neural networks (DNNs), or deep learning, are an extension of the classical ANN that incorporate multiple hidden layers (*65*). Deep learning does not represent a single algorithm, but a broad class of methods with diverse network architectures, including both supervised and unsupervised methods. Deep architectures include multiple processing layers and nonlinear transformations, with the outputs from each layer passed as inputs to the next. Supervised DNNs simultaneously learn a feature representation and a mapping from features to the target, enabling good model performance without requiring well-chosen features as inputs. Ross *et al*. (*66*) provide an illustrative example of a convolutional neural network (CNN), a popular class of DNNs, with convolutional layers for feature extraction and a fully connected layer for classification and regression. However, training a deep network also requires fitting a large number of parameters, which requires large training datasets and techniques to prevent overfitting the model (i.e., memorizing the training data rather than learning a general trend). The complexity of deep learning architectures can also make the models difficult to interpret.

DNNs trained on simulation-generated data can learn a model that approximates the output of physical simulations. DeVries *et al*. (*23*) use a deep, fully connected neural network to learn a compact model that accurately reproduces the time-dependent deformation of Earth as modeled by computationally intensive codes that solve for the response to an earthquake of an elastic layer over an infinite viscoelastic half space. Substantial computational overhead is required to generate simulation data for training the network, but once trained the model acts as a fast

operator, accelerating the computation of new viscoelastic solutions by orders of magnitude. Moseley *et al*. (*67*) use a CNN, trained on synthetic data from a finite difference model, to perform fast full wavefield simulations.

Shoji *et al*. (*20*) use a CNN to classify volcanic ash particles on the basis of their shape, with each of the four classes corresponding to a different physical eruption mechanism. The authors use the class probabilities returned by the network to identify the mixing ratio for ash particles with complex shapes, a task that is difficult for expert analysts.

Several recent studies have applied DNNs with various architectures for automatic earthquake and seismic event detection (*16*, *68*, *69*), phase-picking (*66*, *70*), and classification of volcano-seismic events (*71*). Wiszniowski *et al*. (*72*) introduced a real-time earthquake detection algorithm using a recurrent neural network (RNN), an architecture designed for sequential data. Magaña- Zook and Ruppert (*73*) use a long short-term memory (LSTM) network (*74*), a sophisticated RNN architecture for sequential data, to discriminate natural seismicity from explosions. An advantage of DNNs for earthquake detection is that feature extraction is performed by the network, so minimal preprocessing is required. By contrast, shallow ANNs and other classical learning algorithms require the user to select a set of key discriminative features, and poor feature selection will hurt model performance. Because it may be difficult to define the distinguishing characteristics of earthquake waveforms, the automatic feature extraction of DNNs can improve detection performance, provided large training sets are available.

Araya-Polo *et al*. (*75*) use a DNN to learn an inverse for a basic type of tomography. Rather than using ML to automate or improve individual elements of a standard workflow, they aim to learn to estimate a wave speed model directly from the raw seismic data. The DNN model can compute models faster than traditional methods.

Understanding the fundamental properties and interpretability of DNNs is a very active line of research. A scattering transform (*76*, *77*) can provide natural insights in CNNs relevant to geoscience. This transform is a complex CNN that discards the phase and thus exposes spectral correlations otherwise hidden beneath the phase fluctuations, to define moments. The scattering transform by design has desirable invariants. A scattering representation of stationary processes includes their second-order and higher-order moment descriptors. The scattering transform is effective, for example, in capturing key properties in multifractal analysis (*78*) and stratified continuum percolation relevant to representations of sedimentary processes and transport in porous media, respectively. Interpretable DNN architectures have been obtained through construction from the analysis of inverse problems in the geosciences (*79*); these are potentially large improvements over the original reconstructions and algorithms incorporating sparse data acquisition, and acceleration.

Methods and trends for unsupervised learning

Clustering and self-organizing maps

There are many different clustering algorithms, including k-means, hierarchical clustering, and self-organizing maps (SOMs). A SOM is a type of unsupervised neural network that can be used for either dimensionality reduction or clustering (*80*) [see Roden *et al*. (*81*) for thorough explanation of SOMs]. Carneiro *et al*. (*33*) applied a SOM to airborne geophysical data to identify key geophysical signatures and determine their relationship to rock types for geological mapping in the Brazilian Amazon. Roden *et al*. (*81*) identified geological features from seismic attributes using a combination of PCA for dimensionality reduction followed by SOM for clustering. SOMs are often used to identify seismic facies, but standard SOMs do not account for spatial relationships among the data points. Zhao *et al*. (*82*) propose imposing a stratigraphy constraint on the SOM algorithm to obtain more detailed facies maps. SOMs have also been applied to seismic waveform data for feature selection (*83*) and to cluster signals to identify multiple event types (*84*, *85*).

Supervised and unsupervised techniques are commonly used together in ML workflows. Cracknell *et al*. (*86*) train a RF classifier to identify lithology from geophysical and geochemical survey data. They then apply a SOM to the volcanic units from the RF-generated geologic map to identify subunits that reveal compositional differences. In the geosciences it is common to have large datasets in which only a small subset of the data are labeled. Such cases call for semi-supervised learning methods designed to learn from both labeled and unlabeled data. In a semi-supervised approach, Köhler *et al*. (*87*) detect rockfalls and volcano-tectonic events in continuous waveform data using an SOM for clustering and assigning each cluster a label based on a small number of known examples. Sick *et*

al. (*88*) also use a SOM with nearest-neighbor classification to classify seismic events by type (quarry blast versus seismic) and depth.

Feature learning

Unsupervised feature learning can be used to learn a low-dimensional or sparse feature representation for a dataset. Valentine and Trampert (*32*) learn a compact feature representation for earthquake waveforms using an autoencoder network, a type of unsupervised DNN designed to learn efficient encodings for data. Qian *et al*. (*89*) apply a deep convolutional autoencoder network to prestack seismic data to learn a feature representation that can be used in a clustering algorithm for facies mapping.

Holtzman *et al*. (*27*) use nonnegative matrix factorization and HMMs together to learn features to represent earthquake waveforms. K-means clustering is applied to these features to identify temporal patterns among 46,000 low-magnitude earthquakes in the Geysers geothermal field. The authors observe a correlation between the injection rate and spectral properties of the earthquakes.

Dictionary learning

Sparse dictionary learning is a representation learning method that constructs a sparse representation in the form of a linear combination of basic elements, or atoms, as well as those basic elements themselves. The dictionary of atoms is learned from a set of input data while finding the sparse representations. Dictionary learning methods, which learn an overcomplete basis for sparse representation of data, have been used to de-noise seismic data (*90*, *91*). Bianco and Gerstoft (*92*) develop a linearized (surface-wave) travel-time tomography approach that sparsely models local behaviors of overlapping groups of pixels from a discrete slowness map following a maximum a posteriori (MAP) formulation. They employ iterative thresholding and signed K-means dictionary learning to enhance sparsity of the representation of the slowness estimated from travel-time perturbations.

Deep generative models

Generative models are a class of ML methods that learn joint probability distributions over the dataset. Generative models can be applied to both unsupervised and supervised learning tasks. Recent work has explored applications of deep generative models, in particular generative adversarial networks (GANs) (*93*). A GAN is a system of two neural networks with opposing objectives: a generator network that uses training data to learn a model to generate realistic synthetic data and a discriminator network that learns to distinguish the synthetic data from real training data [see (*94*) for a clear explanation].

Deep generative models, such as the Deep Rendering Model (*95*), Variational Autoencoders (VAEs) (*96*), and GANs (*93*), are hierarchical probabilistic models that explain data at multiple levels of abstraction, and thereby accelerate learning. The power of abstraction in these models allows their higher levels to learn concepts and categories far more rapidly than their lower levels, owing to strong inductive biases and exposure to more data (*97*). The unsupervised learning capability of the deep generative models is particularly attractive to many inverse problems in geophysics where labels are often not available.

The use of neural networks can substantially reduce the computational cost of generating synthetic seismograms compared with numerical simulation models. Krischer and Fichtner (*98*) use a GAN to map seismic source and receiver parameters to synthetic multicomponent seismograms. Mosser *et al*. (*99*) use a domain transfer approach, similar to artistic style transfer, with a deep convolutional GAN (DCGAN) to learn mappings from seismic amplitudes to geological structure and vice versa. The authors' approach enables both forward modeling and fast inversion. GANs have also been applied to geological modeling by Dupont *et al*. (*100*), who infer local geological patterns in fluvial environments from a limited number of rock type observations using a GAN similar to those used for image inpainting. Chan and Elsheikh (*101*) generate realistic, complex geological structures and subsurface flow patterns with a GAN. Veillard *et al*. (*102*) use both a GAN and a VAE (*96*) to interpret geological structures in 3D seismic data.

Other techniques

Reinforcement learning is a ML framework in which the algorithm learns to make decisions to maximize a reward by trial and error. Draelos *et al*. (*103*) propose a reinforcement learning–based approach for dynamic selection of thresholds for single-station earthquake detectors based on the observations at neighboring stations. This approach is a general method for automated parameter tuning that can be used to improve the sensitivity of single-station detectors using information from the seismic network.

Several recent studies in seismology have used techniques for fast near-neighbor search to determine focal mechanisms of seismic events (*104*), to estimate ground motion and source parameters (*105*), or to enable large-scale template matching for earthquake detection (*106*). Each of these three applications requires a database of known or precomputed earthquake features and uses an efficient search algorithm to reduce the computational runtime. By contrast, Yoon *et al*. (*107*) take an unsupervised pattern-mining approach to earthquake detection; the authors use a fast similarity search algorithm to search the continuous waveform data for similar or repeating, allowing the method to discover new events with previously unknown sources. This approach has been extended to multiple stations (*108*) and can process up to 10 years of continuous data (*109*).

Network analysis techniques—methods for analyzing data that can be represented using a graph structure of nodes connected by edges—have also been used for data-driven discovery in the sEg. Riahi and Gerstoft (*110*) detect weak sources in a dense array of seismic sensors using a graph clustering technique. The authors identify sources by computing components of a graph where each sensor is a node and the edges are determined by the array coherence matrix. Aguiar and Beroza (*111*) use PageRank, a popular algorithm for link analysis, to analyze the relationships between waveforms and discover potential low-frequency earthquake (LFE) signals.

Recommendations and opportunities

ML techniques have been applied to a wide range of problems in the sEg; however, their impact is limited (Fig. 5). Data challenges can hinder progress and adoption of the new ML tools; however, adoption of these methods has lagged some other scientific domains with similar data quality issues.

Our recommendations are informed by the characteristics of geoscience datasets that present challenges for standard ML algorithms. Datasets in the sEg represent complex, nonlinear, physical systems that act across a vast range of length and time scales. Many phenomena, such as fluid injection and earthquakes, are strongly nonstationary. The resulting data are complex, with multiresolution, spatial, and temporal structures requiring innovative approaches. Further, much existing data are unlabeled. When available, labels are often highly subjective or biased toward frequent or well-characterized phenomena, limiting the effectiveness of algorithms that rely on training datasets. The quality and completeness of datasets create another challenge as uneven data collection, incomplete datasets, and noisy data are common.

Benchmark datasets

The lack of clear ground-truth and standard benchmarks in solid sEg problems impedes the evaluation of performance in geoscience applications. Ground truth, or reference data for evaluating performance, may be unavailable, incomplete, or biased. Without suitable ground-truth data, validating algorithms, evaluating performance, and adopting best practices are difficult. Automatic earthquake detection provides an illustrative example. New signal processing or ML-based earthquake detection algorithms have been regularly developed and applied over the past several decades. Each method is typically applied to a different dataset, and authors set their own criteria for evaluating performance in the absence of ground truth. This makes it difficult to determine the relative detection performance, advantages, and weaknesses of each method, which prevents the community from adopting and iterating on the best new detection algorithms.

Benchmark datasets and challenge problems have played an important role in driving progress and innovation in ML research. High-quality benchmark datasets have two key benefits: (i) enabling rigorous performance comparisons and (ii) producing better models. Well-known challenge problems include mastering game play (*112*–*115*), competitions for movie recommendation [Netflix prize (*15*)], and image recognition [ImageNet (*14*)]. The performance gain demonstrated by a CNN (*116*) in the 2012 ImageNet competition triggered a wave of research in deep

learning. In computer vision, it is common practice to report performance of new algorithms on standard datasets, such as the MNIST handwritten digit dataset (*117*).

Greater use of benchmark datasets can accelerate progress in applying ML to problems in the sEg. This will require an investment from the research community, both in terms of creating and maintaining datasets and also in reporting algorithm performance on benchmark datasets in published work. The contribution of compiling and sharing benchmark datasets is unlikely to go unrecognized. The ImageNet image recognition dataset (*118*) has been cited in over 6000 papers.

Recently, the Institute of Geophysics at the Chinese Earthquake Administration (CEA) and Alibaba Cloud hosted a data science competition with more than 1000 teams centered around automatic detection and phase picking of aftershocks following the 2008 M_S 8.0 Wenchuan earthquake (*119*, *120*). The ground-truth phase-arrival data, against which entries were assessed, were determined by CEA analysts. Such challenges are useful for researchers seeking to test and improve their detection algorithms. Future competitions should have greater impact if they are accompanied with some form of broader follow-up, such as publications associated with top-performing entries or a summary of effective methods and lessons learned from competition organizers.

Creating benchmark datasets and determining evaluation metrics are challenging when the underlying data are incompletely understood. The ground truth used as a benchmark may suffer from the same biases as training data. Although benchmark datasets can provide a useful guide, the research community must not disregard or penalize methods that discover new phenomena not represented in the ground truth. Performance evaluation needs to include both the overall error rate, along with the relative strengths and weaknesses, including kinds of errors made by the algorithm (*121*). Ideally, within a given problem domain, several diverse benchmark datasets would be available to the research community to avoid an overly narrow focus on algorithm development. For example the performance of earthquake detection algorithms can vary on the basis of the type of events to be detected (e.g., regional events, volcanic tremor, tectonic tremor) and the characteristics of the noise signals. An additional approach would be to create datasets from simulations where the simulated data are released but the underlying model is kept hidden, e.g., a model of complex fault geometry based on seismic reflection data. Researchers could then compete to determine which approaches best recover the input model.

Open science

Adoption of open science principles (*122*) will better position the sEg community to take advantage of the rapid pace of development in AI. This should include a commitment to make code (open source), datasets (open data), and research (open access) publicly available. Open science initiatives are especially important for validating and ensuring reproducibility of results from more-difficult-to-interpret, “black-box” ML models such as DNNs.

Open source codes, often shared through online platforms (*123*, *124*), have already been adopted for general data processing in seismology with the ObsPy (*125*, *126*) and Pyrocko (*127*) toolboxes. Scikit-learn is another example of broadly applied open-source software (*128*). Along with benchmark datasets, greater sharing of the code that implements new ML-based solutions will help accelerate the development and validation of these new approaches. Active research areas like earthquake detection benefit as available open source codes enable direct comparisons of multiple algorithms on the same datasets to assess the relative performance. To the extent possible, this should also extend to the sharing of the original datasets and pretrained ML models.

The use of electronic preprints [e.g., (*129*–*131*)] may also help to accelerate the pace of research (*132*) at the intersection of Earth science and AI. Preprints allow authors to share preliminary results with a wider community and receive feedback in advance of the formal review process. The practice of making research available on preprint servers is common in many science, technology, engineering, and mathematics (STEM) fields, including computer vision and natural language processing—two fields that are driving development in deep learning (*133*); however, this practice has yet to be widely adopted within the sEg community.

New data sources

In recent years new, unconventional data sources have become available but have not been fully exploited, presenting new opportunities for the application and development of new ML-based analysis tools. Data sources such

as light detection and ranging (LiDAR) point clouds (*134*), distributed acoustic sensing with fiber optic cables (*135*–*137*), and crowd-sourced data from smartphones (*17*), social media (*138*, *139*), web traffic (*140*), and microelectromechanical systems (MEMS) accelerometers (*141*) are well-suited to applications using ML. Interferometric synthetic aperture radar (InSAR) data are widely used for applications such as identifying crops or deforestation, but have seen minimal use in ML applications for geological and geophysical problems. High-resolution satellite and multispectral imagery (*142*) provide rich datasets for geological and geophysical applications, including the study of evolving systems such as volcanoes, earthquakes, land-surface change, mapping geology, and soils. A disadvantage of nongovernmental satellite data can be cost. Imagery data from sources such as Google Maps or lower-resolution multispectral data from government-sponsored satellites such as SPOT and ASTER are available without cost.

Machine learning solutions, new models, and architectures

Researchers have many opportunities for collaborative research between the geoscience and ML communities, including new models and algorithms to address data challenges that arise in sEg [see also Karpatne *et al*. (*143*)]. Real-time data collection from geophysical sensors offer new test cases for online learning in streaming data. Domain expertise is required to interpret many geoscientific datasets, making these interesting use cases for the development of interactive ML algorithms, including scientist-in-the-loop systems.

A challenge that comes with entirely data-driven approaches is the need for large quantities of training data, especially for modeling through deep learning. Moreover, ML models may end up replicating the biases in training data, which can arise during data collection or even through the use of specific training datasets. Thus, extracting maximum value from geoscientific datasets will require methods capable of learning with limited, weak, or biased labels. Furthermore, because the phenomena of interest are governed by complex and dynamic physical processes, there is a need for new approaches to analyzing scientific datasets that combine data-driven and physical modeling (*144*).

Much of the representational power of modern ML techniques, such as DNNs, comes from the ability to recognize paths to data inversion outside of the established physical and mathematical frameworks, through nonlinearities that give rise to highly realistic yet nonconvex regularizers. Recently, interpretable DNN architectures were constructed based on the analysis of inverse problems in the geosciences (*79*) that have the potential to mitigate ill-posedness, accelerate reconstruction (after training), and accommodate sparse (constrained) data acquisition. In the framework of linear inverse problems, various imaging operators induce particular network architectures (*26*). Furthermore, deep generative models will play an important role in bridging multilevel regularized iterative techniques in inverse problems with deep learning. In the same context, priors may be learned from the data.

Another approach to mitigating deep learning's reliance on large datasets is to use simulations to generate supplemental synthetic training data. In such cases, domain adaption can be used to correct for differences in the data distribution between real and synthetic data. Domain adaptation architectures, including the mixed-reality generative adversarial networks (*145*), iteratively map simulated data to the space of real data and vice versa. Studying trained deep generative models can reveal insights into the underlying data-generating process, and inverting these models involves inference algorithms that can extract useful representations from the data.

A note of caution in applying ML to geoscience problems: ML should not be applied naïvely to complex geoscience problems. Biased data, mixing training and testing data, overfitting, and improper validation will lead to unreliable results. As elsewhere, working with data scientists will help mitigate these potential issues.

Geoscience curriculum

Data science is moving quickly, and geoscientists would benefit from close collaboration with ML researchers (*146*) to take full advantage of developments at the cutting edge. Collaboration will require focused effort on the part of geoscientists. A greater component of data science and ML in geoscience curricula could help, as would recruiting students trained in data science to work on geoscience research. Interdisciplinary research conferences could and are being used to promote collaborations through identifying common interests and complementary capabilities.

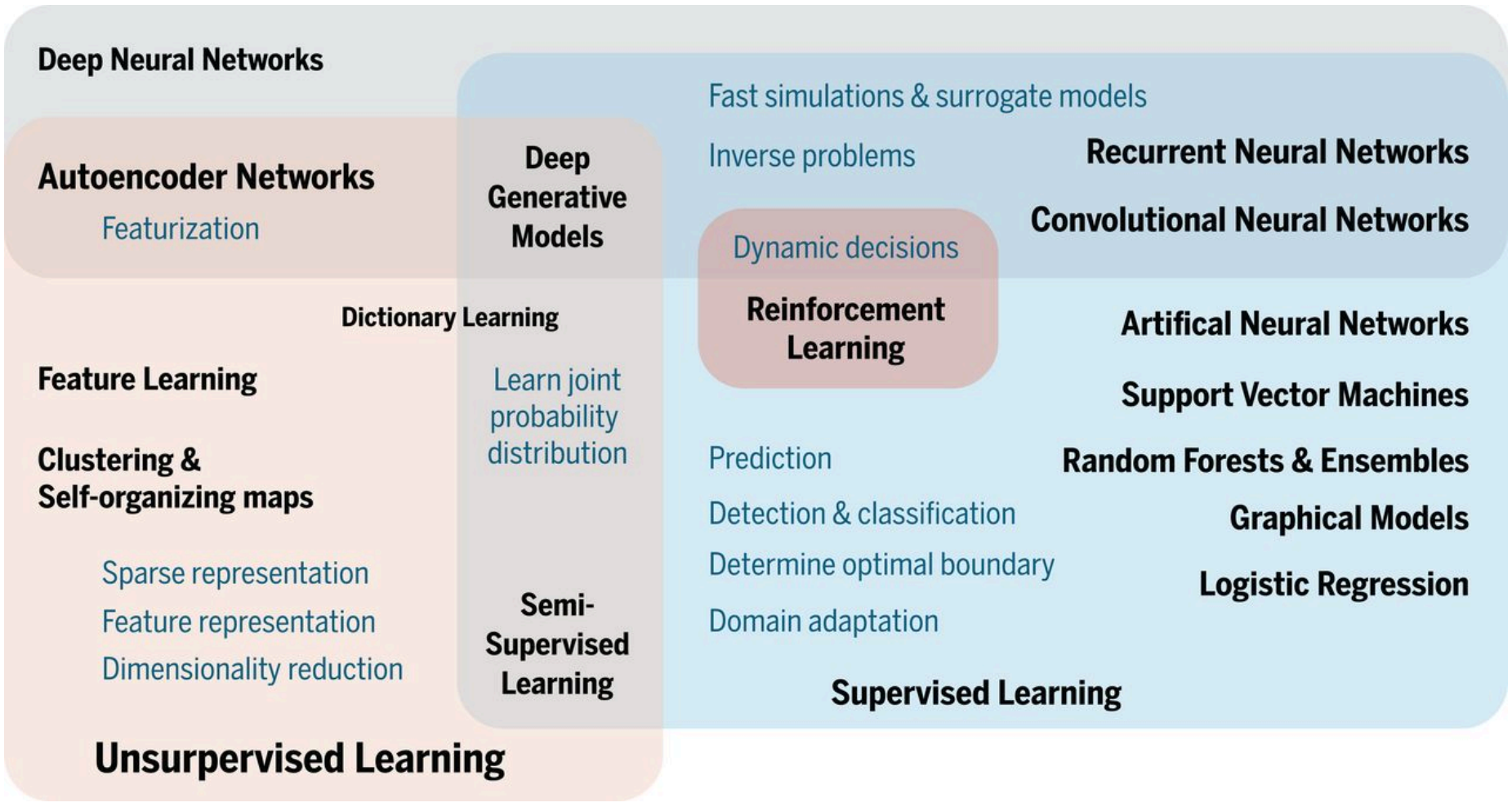


Fig. 4 ML methods and their applications.
Most ML applications in sEg fall within two classes: unsupervised learning and supervised learning. In supervised learning tasks, such as prediction (*21*, *28*) and classification (*16*, *20*), the goal is to learn a general model based on known (labeled) examples of the target pattern. In unsupervised learning tasks, the goal is instead to learn structure in the data, such as sparse or low-dimensional feature representations (*27*). Other classes of ML tasks include semi-supervised learning, in which both labeled and unlabeled data are available to the learning algorithm, and reinforcement learning. Deep neural networks represent a class of ML algorithms that include both supervised and unsupervised tasks. Deep learning algorithms have been used to learn feature representations (*32*, *89*), surrogate models for performing fast simulations (*23*, *75*), and joint probability distributions (*98*, *100*).

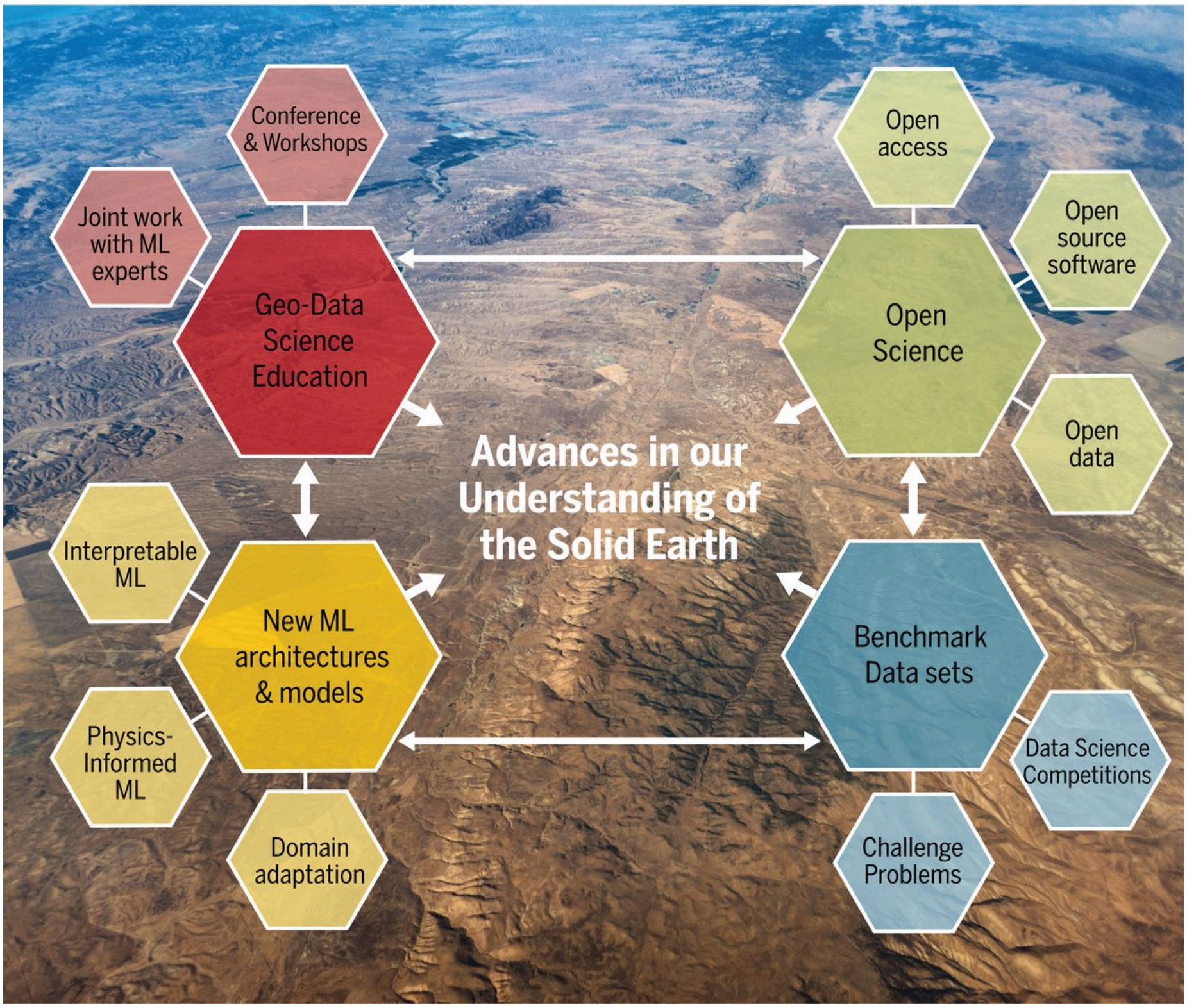


Fig. 5 Recommendations for advancing ML in geoscience.
To make rapid progress in ML applications in geoscience, education beginning at the undergraduate university level is key. ML offers an important new set of tools that will advance science and engineering rapidly if the next generation is well trained in their use. Open-source software such as Sci-kit Learn, TensorFlow, etc. are important components as well, as are open-access publications that all can access for free such as the LANL/Cornell arXiv and PLOS (Public Library of Science). Also important are selecting the proper ML approach and developing new architectures as needs arise. Working with ML experts is the best approach at present, until there are many experts in the geoscience domain. Competitions, conferences, and special sessions could also help drive the field forward.

Acknowledgments

This article evolved from presentations and discussions at the workshop “Information is in the Noise: Signatures of Evolving Fracture Systems” held in March 2018 in Gaithersburg, Maryland. The workshop was sponsored by the Council on Chemical Sciences, Geosciences and Biosciences of the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences. We thank the members of the council for their encouragement and assistance in developing this workshop. **Funding:** K.J.B. and G.C.B. acknowledge support from National Science Foundation (NSF) grant no. EAR-1818579. K.J.B. acknowledges support from the Harvard Data Science Initiative. P.A.J. acknowledges support from Institutional Support (LDRD) at Los Alamos and the Office of Science (OBES) grant KC030206. M.V.d.H. acknowledges support from the Simons Foundation under the MATH + X program and the National Science Foundation (NSF) grant no. DMS-1559587. P.A.J. thanks B. Rouet-LeDuc, I. McBrearty, and C. Hulbert for fundamental insights. **Competing interests:** The authors declare no competing interests.

References and Notes

1 P. Baldi, P. Sadowski, D. Whiteson, Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* **5**, 4308 (2014). 10.1038/ncomms5308

GO TO REFERENCE • CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR

2 A. A. Melnikov, H. P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, H. J. Briegel, Active learning machine learns to create new quantum experiments. *Proc. Natl. Acad. Sci. USA.* **115**, 1221–1226 (2018). 10.1073/pnas.1714936115

CROSSREF • WEB OF SCIENCE • GOOGLE SCHOLAR

eLetters (0)

Ancient alleles drive contemporary climate adaptation in an alpine plant

BY SIMONE FIOR, HIRZI LUQMAN, *ET AL.*

Lithium-ion intercalation by coupled ion-electron transfer

BY YIRUI ZHANG, DIMITRIOS FRAGGEDAKIS, *ET AL.*

ATP-dependent remodeling of chromatin condensates reveals distinct mesoscale outcomes

BY CAMILLE MOORE, EMILY WONG, *ET AL.*

PDF

TABLE OF CONTENTS

LATEST NEWS

NEWS | 6 OCT 2025

Big U.S. West Coast earthquakes could come as a one-two punch

SCIENCEINSIDER | 6 OCT 2025

Medicine Nobel goes to three researchers who identified immune system's security guards

SCIENCEINSIDER | 3 OCT 2025

Science teachers scramble as U.S. climate resources vanish

NEWS | 3 OCT 2025

Fate of the last female great auk is finally solved

SCIENCEINSIDER | 2 OCT 2025

Jane Goodall, famed primatologist, changed the way we thought about apes

SCIENCEINSIDER | 2 OCT 2025

Renowned U.S. climate center trims staff ahead of expected budget cuts