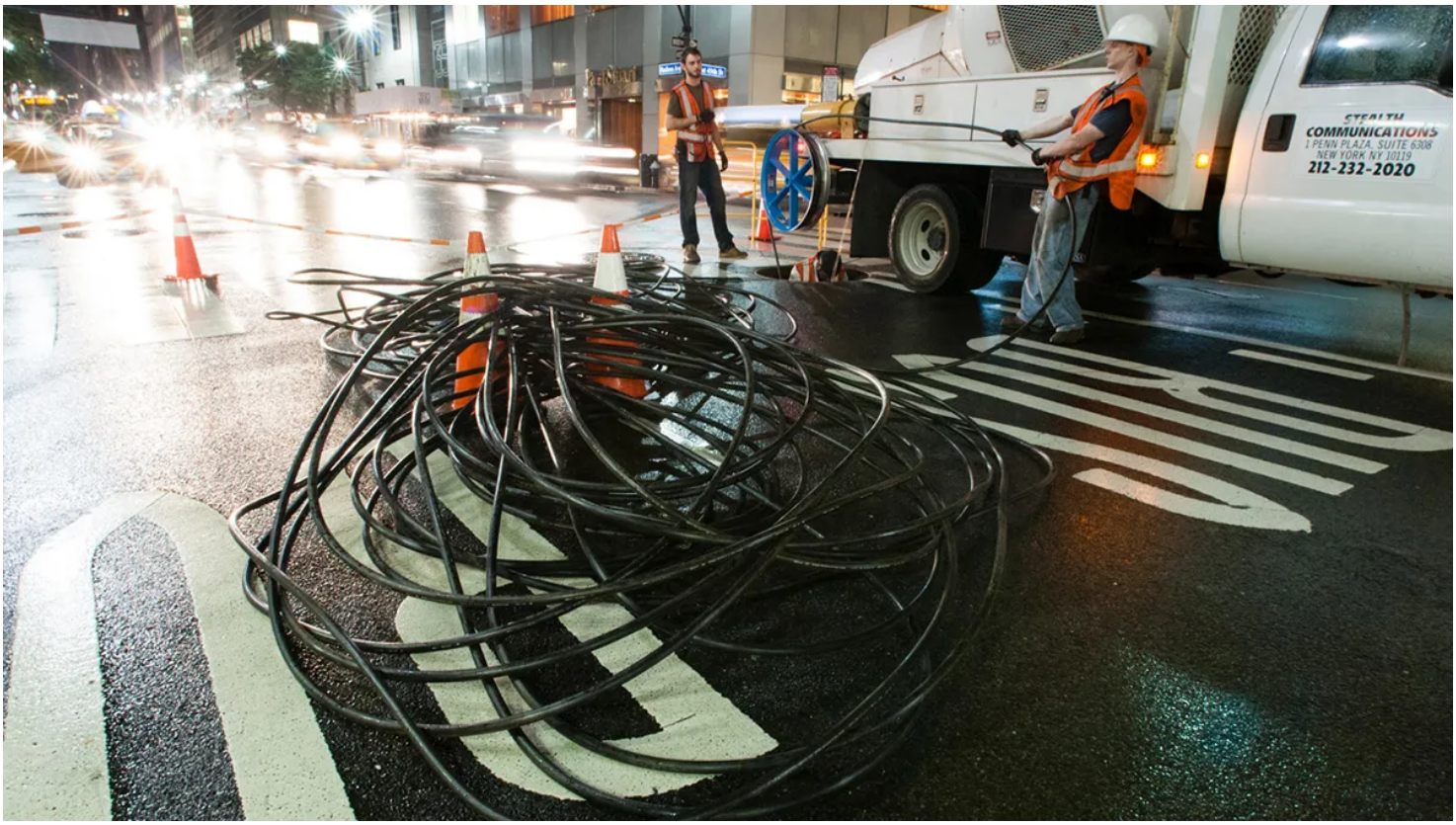
Eos

Distributed Sensing and Machine Learning Hone Seismic Listening

Fiber-optic cables can provide a wealth of detailed data on subsurface vibrations from a wide range of sources. Machine learning offers a means to make sense of it all.

By **Whitney Trainor-Guitton**, **Eileen R. Martin**, **Verónica Rodríguez Tribaldos**, **Nicole Taverna**, and **Vincent Dumont**
4 March 2022



A work crew installs fiber-optic telecommunications cable under a street in midtown Manhattan in New York City. This type of cable can function as sensitive seismic strain sensors, providing Earth scientists with a deluge of data. Credit: Stealth Communications/Wikimedia Commons, CC BY-SA 3.0

Finding solutions to several of the most pressing environmental, energy, and geohazards challenges of our time—from carbon sequestration to major earthquakes—depends on our ability to understand Earth's subsurface and how we interact with it. This understanding largely emerges from the study of underground vibrations that result from both natural events (e.g., earthquakes, magma movements, and even strong winds and rainstorms) and human activities (e.g., hydraulic fracturing, excavation, and vehicular traffic).

Seismometer networks are conventionally used to map and track these vibrations, but setting up these networks can be difficult, expensive, or impossible, depending on the setting. A recent alternative is to gather data on vibrations from thousands of kilometers of underground fiber-optic cables already in place for telecommunications networks. However, sifting through the deluge of data these networks produce is an unwieldy task. Machine learning can help make it less daunting.

A New Use for Fiber-Optic Cables

Seismological studies, which locate and measure vibrations underground, provide a wealth of information on subsurface structure and dynamic surface–subsurface interactions. For these data to be fully useful and comparable among various studies, however, it is crucial that they are acquired at the right places and times and at appropriate spatial and temporal resolutions.

> *For seismic data to be fully useful and comparable among various studies, it is crucial that they are acquired at the right places and times and at appropriate spatial and temporal resolutions.*

In many contexts, using traditional seismic "point" sensors to achieve these requirements can be challenging. For example, broadband seismometers that record seismic vibrations—including earthquakes, **induced seismicity**, seismic waves generated intentionally for subsurface imaging, and other natural and human-caused vibrations—are often expensive to purchase and costly and labor intensive to deploy. Furthermore, installing them—especially in remote environments or urban areas—often involves logistical and legal considerations related to data acquisition budgets, land access, or permitting, so many studies end up having limited temporal and spatial coverage.

In the past half decade, distributed acoustic sensing (DAS) **has emerged** as an attractive and versatile alternative to conventional seismometer deployments (Figure 1). DAS technology turns conventional fiber-optic cables into arrays of hundreds to thousands of **strain rate sensors** that enable continuous, long-term data acquisition over distances of tens of kilometers, with spatial resolutions on the order of a few meters, and at frequencies ranging from millihertz to kilohertz, allowing for detection of all manners of seismic and other types of vibrations. These vibrations momentarily distort the shape of a fiber and cause changes in the travel time of light pulses scattered throughout the fiber. Instrumentation installed at one end of the fiber, called an interrogator, can detect these changes in travel time and convert these data into information about the location of the shaking.

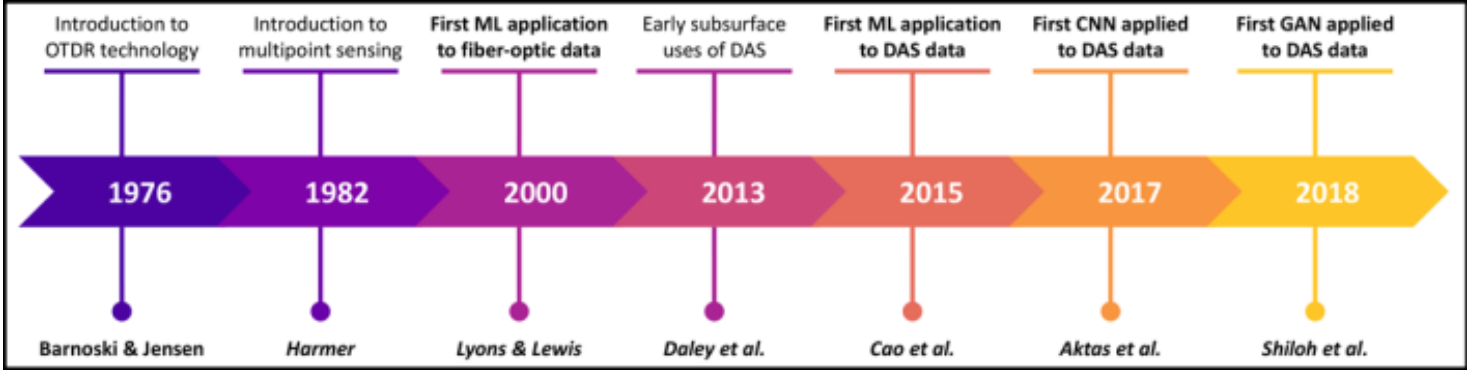


Fig. 1. This timeline shows significant experimental and machine learning (ML) developments with respect to fiber-optic data, beginning with the invention of optical time domain reflectometer (OTDR) technology, which allows characterization of the condition and light transmission performance of fiber-optic cables and is the measurement principle used in most modern to distributed acoustic sensing (DAS) systems. More recently, a convolutional neural network (CNN) and a generative adversarial network (GAN)—a deep learning approach and an ML approach, respectively—have been applied DAS data. Seminal studies describing each development are indicated below the dates: Barnoski and Jensen, 1976, https://doi.org/10.1364/AO.15.002112; Harmer, 1982, https://doi.org/10.1177/002029408201500403; Lyons and Lewis, 2000, https://doi.org/10.1177/014233120002200504; Daley et al., 2013, https://doi.org/10.1190/tle32060699.1; Cao et al., 2015, https://doi.org/10.1364/ACPC.2015.ASu2A.145; Aktas et al., 2017, https://doi.org/10.1117/12.2262108; Shiloh et al., 2018, https://doi.org/10.1364/OFS.2018.ThE22.

DAS can provide both scientific and economic advantages over point sensors in situations in which fiber-optic cables are already installed, such as in the case of existing fiber-optic telecommunications networks. Unused, or "dark," fibers in these networks can be repurposed as sensing arrays. If not already available, fiber can often be installed or trenched relatively easily, especially in comparison to the amount of time and effort required to deploy the same number of point sensors. And unlike geophones or seismometers, fiber-optic cables can be left to collect data for long periods of time without being disturbed by harsh environmental conditions, wildlife, or people.

Telecommunications grade fiber is suitable for shallow borehole and surface deployments and can cost only a few dollars per meter. Furthermore, having a permanently deployed receiver array is advantageous for time-lapse DAS surveys and eliminates the need for costly instrument servicing. The interrogator, the most expensive part of the measurement system, can be used on site only when measurements are required to further reduce expenses. The cost of DAS interrogators can still be high for extended duration deployments, but achieving equivalent coverage and spatial density of measurements with conventional sensors is currently prohibitive. Moreover, we expect the costs of DAS to continue declining. Thus, when compared with large-scale deployments of point sensors with lower coverage and density, the scientific value of DAS systems may justify the costs.

Recent work has validated DAS's usefulness across numerous applications, giving scientists unprecedented views into many targets of interest. For example, research has demonstrated that DAS can act as a permanent seismic sensor for monitoring carbon dioxide movement through a storage reservoir [e.g., ***Daley et al.***, 2013], a finding relevant for potential carbon sequestration efforts. Other studies have used DAS to measure dynamic strain in **volcanic environments** [***Jousset et al.***, 2018], to characterize physical properties of near-surface rock structure [***Ajo-Franklin et al.***, 2019], to monitor icequakes and other events related to glacier

movement and dynamics [*Walter et al.*, 2020], and to detect regional earthquakes and ocean wave dynamics using existing cables on the seafloor [e.g., *Sladen et al.*, 2019].

> *Challenges in storing, managing, exploring, and analyzing vast amounts of data collected with distributed acoustic sensing (DAS) are hindering its wider application, resulting in heaps of rich data that remain unused.*

However, challenges in storing, managing, exploring, and analyzing the vast amounts of data collected by DAS are hindering its wider application, resulting in heaps of rich data that remain unused. Consider the example of a researcher who spends entire 40-hour work weeks combing through a city-scale DAS data set of measurements from 10,000 sensors at a realistic pace. If this person looked at 250 sensors at a time, examining 30-second windows of recorded data for 15 seconds apiece, perhaps to label vibrational events of interest, roughly 84 weeks would be required to completely analyze just 1 week (typically at least several terabytes) of recorded data.

Although cumbersome to analyze manually, these dense data are increasingly amenable to the application of traditional and innovative **machine learning** (ML), a set of highly flexible tools that use algorithms to parse and learn from data and then apply this knowledge to make predictions. Thus, ML is allowing scientists to learn from DAS data more efficiently and to develop transformative techniques for subsurface exploration.

Applying Machine Learning to DAS

ML algorithms can handle very large volumes of data, enabling fast and efficient processing and interpretation of variable and complex observables. With recent and rapid growth in the quantity and variety of DAS data sets, scientists have been exploring how ML techniques can be used to process or identify patterns in DAS data in comparison with classic approaches. In contrast to sparse seismometer networks, DAS records data along evenly spaced cable segments, making it more natural to organize the data into a matrix or image. Much as we do in **computer vision**, we often break up these large matrices into smaller matrices, each representing a collection of neighboring sensors during a short time window.

Since its first application in the early 2000s, DAS+ML, as we abbreviate this combination, has been applied in both **aboveground** and **subsurface** studies. The principal motivation for using DAS+ML to date has been surveillance—monitoring movement and activity to ensure the integrity of pipelines and other infrastructure and the security of perimeters of sensitive locales like airports (Figure 2, top). Researchers turned to ML algorithms to discriminate between the different sources that could generate shaking—did a pedestrian, car, or excavator cause the vibration?

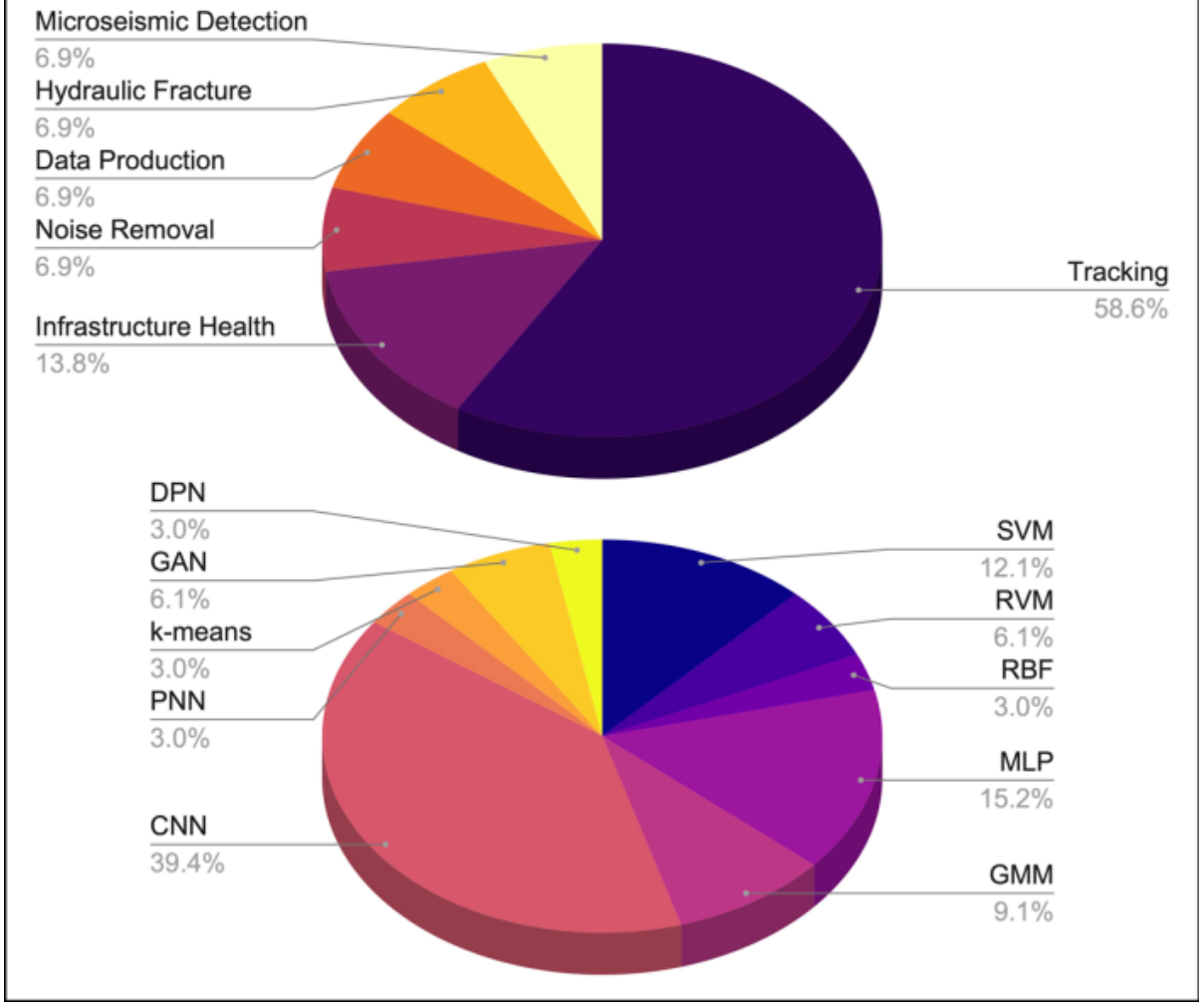


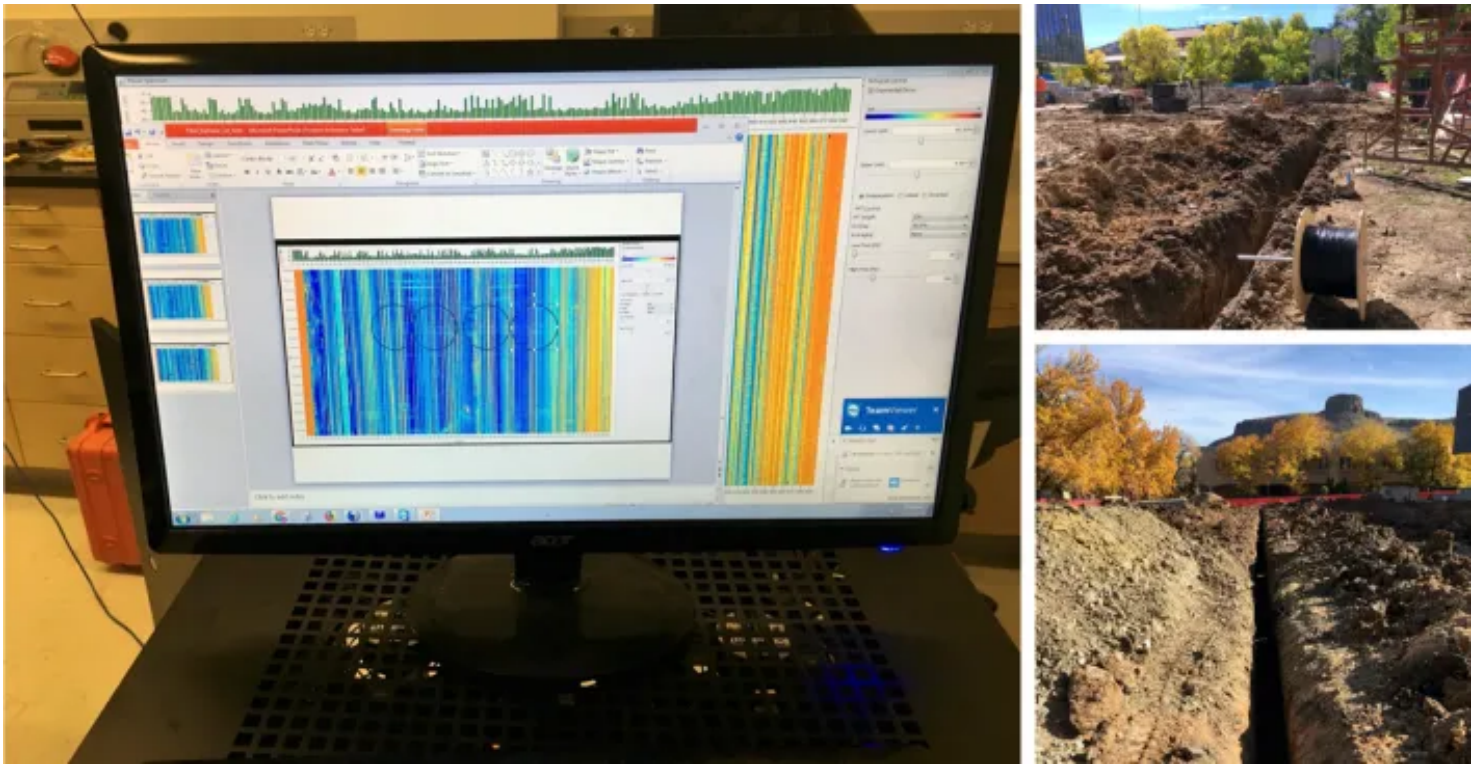
Fig. 2. Applications of DAS technology in geoscience to date (top) have included tracking the movement of vehicles, tractors, and trains; assessing infrastructure health; detecting microseismicity; monitoring hydraulic fracturing; producing data (i.e., generating a large database of tagged events); and removing noise from data sets. ML approaches used on DAS data so far (bottom) include the following: CNN, convolutional neural network; MLP, multilayer perceptron; SVM, support vector machine; GMM, Gaussian mixture model; GAN, generative adversarial network; RVM, relevance vector machine; DPN, dual-path network; k-means, a clustering algorithm; PNN, probabilistic neural network; RBF, radial basis function. Percentages refer to the fraction of papers from our comprehensive list of ML applications to DAS data that cover each application or approach.

Different types of machine learning algorithms have been used for data analysis (Figure 2, bottom), and they have evolved over time. Early studies of distributed vibration sensing used methods like nearest neighbor and support vector machines that required practitioners to determine manually which features of a data set could be detected and discriminated [*Tejedor et al.*, 2017].

More recently, the literature shows a clear shift toward processing and classification of DAS data via deep learning methods, representing the many flavors of convolutional neural networks [*Shao et al.*, 2020]. Unlike basic machine learning models, which need significant human guidance to ensure that they make physically realistic predictions, a deep learning algorithm can determine on its own whether a prediction is plausible through its own neural network. Such work is applying DAS+ML in ever more nuanced ways, with practitioners using it as a diagnostic to assess the "health" of infrastructure, such as tracking moving trains precisely and monitoring their maintenance [*Kowarik et al.*, 2020].

Subsurface investigations are also becoming popular targets for DAS+ML. In these studies, the focus shifts from monitoring infrastructure to **detecting earthquakes** among a noisy mix of vibrations from various other sources or to identifying ambient seismic noise signals that may be useful in imaging or monitoring underground structure or processes. DAS recordings in studies of the subsurface tend to be spatially or temporally large (or both) and contain signals from a wide variety of natural and anthropogenic sources. These factors complicate data labeling, the process of adding contextual information to a data set—for example, to link a particular seismic signal to a recent earthquake—to help an ML algorithm learn. Thus, such studies commonly use unsupervised ML methods, in which algorithms don't require training with **labeled** data sets and can find patterns and evaluate data on their own with little or no human intervention [***Martin et al.***, 2018].

> *Practitioners are applying DAS and machine learning (ML) together in ever more nuanced ways, including using it as a diagnostic to assess the "health" of infrastructure and for subsurface investigations.*



Trenches for a DAS array at Kafadar Commons on the campus of Colorado School of Mines (right) were dug during construction of a nearby building in October 2017. A DAS interrogator was connected to the array in November 2018, measuring strain along the cable in time (y-axis) and space (x-axis) as seen on a computer monitor (left). Credit: Whitney Trainor-Guitton

For down-well microseismic observations (those collected in wells or boreholes), labeling data with certainty is nearly impossible because the seismic sources may be weak, meaning the signal generated does not stand out clearly from background observations. Also, the sources are not directly observed: We don't see the small fractures opening in the subsurface that create a microseismic event. So when supervised ML is used, researchers rely on

supplementing training data sets with **synthetic data** (e.g., data sets generated by wave propagation simulations) to identify microseismic events within the DAS records.

High-Performance, Cloud, and Edge Computing

DAS surveys often involve data collected at rates greater than 10 megabytes per second, producing data sets on the order of 1 terabyte per day [***Shiloh et al.***, 2019]. Processing such quantities of data is computationally expensive and often unfeasible without high-performance computing (HPC) capabilities. Fortunately, significant breakthroughs have been made in both ML and data-intensive HPC, concurrent with the development of DAS technology.

Two developments in particular are allowing geophysicists to significantly speed up ML algorithm training with large DAS data sets. The first is access to highly scalable distributed systems—supercomputers in which components installed on several machines can be combined and adapted to optimize performance and accommodate growth. The second is the development of accelerated computing platforms, including Compute Unified Device Architecture (**CUDA**) and other interfaces for computing with graphics processing units or tensor processing units.

Storage of large volumes of data poses additional challenges commonly faced in DAS investigations. **Cloud storage** can help to address these challenges by providing secure, reliable, and almost infinitely scalable storage solutions. Cloud storage also helps to streamline public accessibility by making data sets readily available from anywhere. Providing such accessibility to valuable data sets is crucial for research transparency, reproducibility, and fostering collaborations among scientists, which ultimately accelerates technological advancement.

> *Transferring data from where they are produced to be stored elsewhere before they are analyzed can be challenging and preclude their use in real-time decisionmaking.*

Cloud computing services paired with cloud storage and data lakes (repositories of data in their native formats) further allow users to work with data where they are stored rather than requiring users to download the data first, eliminating the need for costly data transfers. Storage and analysis on cloud systems have the potential to increase the efficiency of DAS data processing by allowing multiple users to do this processing simultaneously and collaboratively.

For experiments being conducted in remote or poorly connected locations, however, transferring data from where they are produced to be stored elsewhere—whether in the cloud or in localized storage—before they are analyzed can sometimes be challenging and preclude their use in real-time decisionmaking, as data may

reach the data analysis location after their window of peak usability. This limitation is becoming increasingly relevant as the applications of DAS broaden. For example, DAS data collected from within a well during a hydraulic fracturing experiment can aid in determining when and where fractures occur—useful information for monitoring whether injection or production parameters such as flow rates or proppant concentrations should be changed. The bandwidth for data transmission during downhole experiments is often limited, though, making it difficult to extract and process the data fast enough to use them to make quick decisions about adjusting experimental stimulation parameters.

Edge computing, in which data storage and computational resources are located close to where data are produced or consumed rather than in a centralized (but often distant) location, offers a solution. This method involves processing data sets "on the edge" alongside sensors as the data are collected and extracting only useful information from the massive data streams. Edge computing reduces the amount of data that must be transferred. This capability could allow enhanced and automated interconnectivity between DAS, other geophysical or environmental sensors in an area, and researchers' computers through an "Internet of things" platform. Existing edge computing algorithms for distributed sensing generally lack robustness for use across different applications because they are created on a case-by-case basis. However, this approach could be improved by designing ML tools that run efficiently on the edge.

The Limitless Future of DAS+ML

DAS is becoming a transformative tool for studying subsurface processes, but its large data volumes obstruct even wider use. Important design improvements in ML algorithms for DAS data handling and analysis will facilitate the technique's broader application in addressing important scientific questions and societal needs. These improvements involve solutions enabling fast and accurate labeling of training data sets for improved use in supervised algorithms; creating domain-specific features for preparing and choosing characteristics of DAS data (e.g., to distinguish the signal of an earthquake from that of a passing car or train); and improving physics-based pattern discovery in data.

Further innovation is also needed for DAS+ML to be adopted as a tool for initial quality control of seismic data—for example, to identify whether there are spatial or temporal sections of a data set that are of higher quality. Identifying sections of data with the desired frequency content, fiber coupling, or characteristic signals offers more potential for interpreting and distinguishing geologic units correctly, identifying earthquakes amid background noise, discerning natural from human-made seismic events, identifying different types of vehicles for traffic tracking purposes, and other applications. Improving the efficiency of seismic data processing will lead to automated interpretation of DAS data.

As the interpretive abilities of ML algorithms improve, more advanced seismological problems beyond data quality control or signal classification may be

tackled. Such problems include data inversion (estimating maps of subsurface properties from observed data), event forecasting, and uncertainty quantification. In addition, because DAS is unlikely to be used in isolation in the future, necessary advances include ML models that can integrate DAS with other seismic sensor networks and that can transfer and apply models trained on DAS data to data gathered by other types of sensors (or vice versa).

“Necessary advances include ML models that can integrate DAS with other seismic sensor networks and that can transfer and apply models trained on DAS data to data gathered by other types of sensors.

DAS’s increased spatial coverage and density compared with conventional point sensors, when combined with ML, present some exciting benefits, including for improved groundwater resource monitoring and management at local to regional scales, real-time earthquake detection, and rapid deployment of seismic monitoring arrays. The technique should also enable emerging strategies to mitigate climate change: With its usefulness in studying subsurface structure, for example, DAS+ML could help design energy-efficient infrastructure, harness geothermal energy, and investigate options for carbon dioxide sequestration and storage. We see DAS+ML as a critical component in pushing the use of dense seismic data beyond basic research and in providing decisionmakers with the interpretable and actionable results they need to make progress toward these societally important advances.

Acknowledgments

The authors are part of the Working Group on DAS and Machine Learning, which is part of the Distributed Acoustic Sensing Research Coordination Network (RCN) supported by National Science Foundation award 1948737 under the Geosciences and Engineering directorate. We invite readers working in this area to get involved in the RCN. We especially thank DAS+ML RCN working group member Bin Luo for insightful conversations during the development of this article.

References

Ajo-Franklin, J. B., et al. (2019), Distributed acoustic sensing using dark fiber for near-surface characterization and broadband seismic event detection, *Sci. Rep.*, *9*, 1328, https://doi.org/10.1038/s41598-018-36675-8.

Daley, T. M., et al. (2013), Field testing of fiber-optic distributed acoustic sensing (DAS) for subsurface seismic monitoring, *Leading Edge*, *32*(6), 699, https://doi.org/10.1190/tle32060699.1.

Jousset, P., et al. (2018), Dynamic strain determination using fibre-optic cables allows imaging of seismological and structural features, *Nat. Commun.*, *9*, 2509, https://doi.org/10.1038/s41467-018-04860-y.

Kowarik, S., et al. (2020), Fiber optic train monitoring with distributed acoustic sensing: Conventional and neural network data analysis, *Sensors*, *20*(2), 450, https://doi.org/10.3390/s20020450.

Martin, E. R., et al. (2018), A seismic shift in scalable acquisition demands new processing: Fiber-optic seismic signal retrieval in urban areas with unsupervised learning for coherent noise removal, *IEEE Signal Process. Mag.*, *35*(2), 31–40, https://doi.org/10.1109/MSP.2017.2783381.

Shao, L. Y., et al. (2020), Data-driven distributed optical vibration sensors: A review, *IEEE Sens. J.*, *20*, 6,224–6,239, https://doi.org/10.1109/JSEN.2019.2939486.

Shiloh, L., et al. (2019), Efficient processing of distributed acoustic sensing data using a deep learning approach, *J. Lightwave Technol.*, *37*(18), 4,755–4,762, https://doi.org/10.1109/JLT.2019.2919713.

Sladen, A., et al. (2019), Distributed sensing of earthquakes and ocean-solid Earth interactions on seafloor telecom cables, *Nat. Commun.*, *10*, 5777, https://doi.org/10.1038/s41467-019-13793-z.

Tejedor, J., et al. (2017), Machine learning methods for pipeline surveillance systems based on distributed acoustic sensing: A review, *Appl. Sci.*, *7*, 1–26, https://doi.org/10.3390/app7080841.

Walter, F., et al. (2020), Distributed acoustic sensing of microseismic sources and wave propagation in glaciated terrain, *Nat. Commun.*, *11*, 2436, https://doi.org/10.1038/s41467-020-15824-6.

Author Information

Whitney Trainor-Guitton, Colorado School of Mines, Golden; also at Zanskar Geothermal, Provo, Utah; Eileen R. Martin (eileenrmartin@vt.edu), Virginia Polytechnic Institute and State University, Blacksburg; Verónica Rodríguez Tribaldos, Lawrence Berkeley National Laboratory, Berkeley, Calif.; Nicole Taverna, Colorado School of Mines, Golden; and Vincent Dumont, Lawrence Berkeley National Laboratory, Berkeley, Calif.

Citation: Trainor-Guitton, W., E. R. Martin, V. Rodríguez Tribaldos, N. Taverna, and V. Dumont (2022), Distributed sensing and machine learning hone seismic listening, *Eos, 103,* https://doi.org/10.1029/2022EO220121. Published on 4 March 2022.

2022. The authors. CC BY 3.0
Except where otherwise noted, images are subject to copyright. Any reuse without express permission from the copyright owner is prohibited.

© 2022 American Geophysical Union. All rights reserved.