Carnegie Mellon University
Wilton E. Scott Institute
for Energy Innovation

NATIONAL
ENERGY
TECHNOLOGY
LABORATORY

REAL-TIME DECISION-MAKING FOR THE SUBSURFACE REPORT

June 2019

Disclaimer

This report was prepared as an account of work sponsored by Carnegie Mellon University and an agency of the United States Government. Neither Carnegie Mellon University, the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference therein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University, the United States Government, or any agency thereof. The views and opinions of authors expressed therein do not necessarily state or reflect those of Carnegie Mellon University, the United States Government, or any agency thereof.

Table of Contents

Figures

Tables

Authors

Giorgia Bettin
Geothermal Research Manager, Sandia National Laboratory

Grant Bromhal
Senior Fellow, Geologic and Environmental Systems, National Energy Technology Laboratory

Mike Brudzinski
Professor, Miami University

Alan Cohen
Director, Office of Oil and Natural Gas Research, Department of Energy’s Office of Fossil Energy

George Guthrie
Technical Project Manager, Los Alamos National Laboratory

Paul Johnson
Senior Fellow and Technical Staff Member, Los Alamos National Laboratory

Lewis Matthews
Data Scientist, CrownQuest Operating, LLC

Srikanta Mishra
Senior Research Leader, Battelle

Derek Vikara
Subsurface Analysis Program Manager, KeyLogic Systems, Inc.

Editor

Grant Bromhal
Senior Fellow, Geologic and Environmental Systems, National Energy Technology Laboratory

Acronyms and Abbreviations

AAE	Average absolute error
ANN	Artificial neural network
CCUS	Carbon capture, utilization, and storage
DA	Data analytics
DAS	Distributed acoustic sensing
DM	Data management
DOE	Department of Energy
EOR	Enhanced oil recovery
FAIR	Findable, accessible, interoperable, and reusable
FMD	Frequency magnitude distribution
GBM	Gradient Boosting Machine
GR	Gamma ray
InSAR	Interferometric synthetic aperture radar
ML	Machine learning
MSE	Mean squared error
NDA	Non-disclosure agreement
OLS	Ordinary least squares
OPAT	One-parameter-at-a-time
R&D	Research and development
RF	Random forest
ROM	Reduced-order model
RMSE	Root mean square error
SME	Subject matter expert
SURE	Scale, uncertainty, resolution, environment/economics
SVM	Support Vector Machine
SVR	Support Vector Regression
TLP	Traffic Light Protocol
U.S.	United States

Executive Summary

What Were the Workshop Goals?

This report summarizes findings from the “Real-Time Decision-Making for the Subsurface” workshop hosted by Carnegie Mellon University's Wilton E. Scott Institute for Energy Innovation in July 2018. Several dozen technical experts from industry, universities, national laboratories, and the Department of Energy (DOE) convened for two days to discuss the current state of technology that could enable autonomous monitoring and control of subsurface systems, to identify potential short-term and long-term opportunities, and to discuss gaps and challenges that need to be overcome to make this transformational technology a reality. Two primary technology applications were covered at the workshop: unconventional oil and gas recovery and carbon storage. However, the approaches and techniques discussed have utility for a broad range of subsurface activities, such as geothermal energy, subsurface energy storage, and enhanced oil recovery (EOR).

The timeliness of this effort is supported by several recent technological advances. Machine learning (ML), data analytics (DA), and data management (DM) have expanded rapidly in many commercial sectors, providing an array of resources that can be leveraged for subsurface applications. Novel sensors can now provide temporal and spatial resolution not available a decade ago, and technologies under development now may provide the ability to revolutionize how the subsurface is imaged in the coming decade. Developments in sophisticated physics-based simulators, coupled with high performance computing capabilities, enable better prediction of subsurface systems than ever before. Finally, ML, computational speed, and the ability to handle very large data streams have markedly advanced. In short, it is the ideal time to pursue the development of real-time decision-making capabilities that could transform approaches used to develop subsurface energy systems.

What is Real-Time Decision-Making?

Real-time decision-making applicable for engineered subsurface energy systems, which is closely tied to autonomous control of subsurface systems, is a long-term, transformational goal that is likely to take a decade or more to achieve in any broad way. Nonetheless, several technologies that can enable better decision-making and lead to increased productivity and reduced risk by leveraging potentially large and disparate datasets are ready for development in the near-term. Collectively, these approaches have the potential to completely change the way that oil and gas and other subsurface fields are operated. Higher resolution sensors and faster computational platforms are enabling technologies. Novel DM and analytics approaches must be developed to ensure that such advances can be optimized. In light of this, several topics were covered during the workshop that have both near-term and long-term applications, including the identification of

1) use cases for technologies related to ML, DA, and DM, including current and near-term applications as well as transformational opportunities;

2) data types and opportunities to use and strengthen existing datasets;

3) potential DA and ML approaches; and

4) research and development (R&D) challenges.

Because of the broad applicability of these tools and techniques to many aspects of carbon storage operations and unconventional oil and gas exploration and recovery, encompassing surface, near-surface, and deep subsurface applications, the workshop participants identified a handful of priority use cases with the greatest potential benefits. These high-value use cases include

- **Safety:** Protecting workers, the public, the environment, and equipment from potentially significant incidents through technologies that can improve early warning for potential failures, as well as augment or provide context for the information that decision-makers must use when making rapid-fire decisions regarding such incidents
- **Drilling/Geo-steering:** Enhancing resource utilization by speeding up and otherwise improving the drilling process, as well as by improving the placement of wells within the targeted formation
- **Completions and Well Operations Optimization:** Improving recovery and storage efficiencies with next-generation completion design; optimizing injection and production rates; and managing stress states, fracture formation, and fracture growth
- **Reservoir Management/Operations:** Optimizing recovery and storage efficiencies by improving reservoir characterization; designing strategies for more efficient and longer-term approaches to injection and production; ensuring no leakage of the resource outside of the reservoir; and limiting induced seismicity

Making advances in these areas will lead to the achievement of significant progress in unconventional oil and gas recovery, such as significantly improved hydrocarbon recovery efficiency from domestic resources, and in carbon storage, through development of the resources needed to demonstrate safe storage of injected fluids to non-operators and the broader public. More detail on each of these use cases can be found in the body and appendices of this report.

Given the data-driven nature of the approaches being discussed, several conversations at the workshop focused around data availability, data needs, and data quality, as well as the ML and DA tools and techniques that can be used to make the most of this valuable information. The body of this report describes several tools and techniques that can be used to analyze the types of data that can be collected by operators, service companies, and contractors in the field to best facilitate analyses that could generate improvements aligned to the use cases outlined above. The breakout session summaries in the appendices also cover many of the data types needed for the different use cases. Publicly-available datasets, as well as those curated by vendors (i.e., DrillingInfo and IHS Markit), are one potential avenue where readily-available data could expedite foundational analyses. However, publicly available datasets alone are often considered insufficient to achieve targeted solutions to all the issues addressed here, but such data often exist and can help augment site-specific field data. Finding and acquiring this data are also sometimes difficult; creating tools that would enable potential users to search for and identify such publicly available resources would improve the ability to use "big data" (i.e., advanced analysis of very large, diverse datasets) approaches for developing data-driven insights.

Data analytics (DA) is a process used to identify hidden patterns and relationships in large, complex, and multivariate datasets. The application of data analytics involves three key elements:

- ***Data organization and management***: ensuring that the right data is collected, retrieved, and stored for given analyses
- ***Analytics and knowledge discovery***: software-driven modeling to capture input-output relationships
- ***Decision support and visualization***: sharing results with decision-makers, as well as streamlining repetitive tasks

Machine learning (ML) refers to a set of algorithms, based on advanced statistics and computer science for model building, that underpin the analytics and knowledge discovery aspects of DA.

What are Some Key Challenges?

The workshop participants identified several key challenges or barriers that could inhibit the development and use of real-time decision-making tools for subsurface applications. The most notable challenges include

- **Missing or incomplete datasets:** In many cases, the data that are collected by operators is insufficient to perform the analyses needed to make significant improvements in operational efficiencies. Because of cost considerations, operators typically collect the bare minimum amount of data required to sustain operations, or data required to satisfy regulatory reporting requirements; for some of the same reasons, the data quality is also often lacking or datasets are absent from important parameters that were not recorded/collected that may provide useful insights.
- **Incompatibility of data formats:** Proprietary data formats are common across the oil and gas and carbon storage industries. Because of this, when multiple operators, service companies, or vendors work together on the same site, there are significant barriers to common use of the data that are collected. Additionally, data from disparate sources, including operators working in the same field, could be out of sync in terms of timestamp, format, completeness, etc. As a result, data interpretation, preprocessing, and potential importation of missing values becomes a substantial task.
- **Limitations on storing, transmitting, and managing large data volumes:** As novel sensing technologies have evolved and volumes and rates of data that are collected

continue to grow exponentially, the computational tools required to store and transmit such data volumes in real time have often not kept pace, such that some data that are currently collected are not used in a meaningful way. Methods for keeping track of data as it moves through a system that allows storage, backup, and analysis are lacking.

- **Lack of appropriate or useful signals:** In some cases, it is unknown whether signals exist within the collected data or not. Particularly, data that have often been categorized as "noise" (i.e., corrupted, distorted, or otherwise meaningless data) can sometimes contain useful information but because of prior classification as noise, it is either discarded, not collected, or not analyzed.
- **Hesitance to adopt new technologies by decision-makers:** Cultural biases based on personal experience can cause decision-makers in some organizations to not adopt ML and DA approaches to improve their organizational performance.
- **Limitations in labeled datasets for ML applications:** Labeled datasets are required for supervised ML. Often geophysical and geoscience datasets are not labeled or poorly labeled. An example is an earthquake catalog—some events may be mis-identified, seismic wave phases may be mis-picked, etc. Many geoscience datasets suffer from similar problems, leading to errors in the supervised learning procedure.

More information regarding these and other challenges or barriers can be found in the body of the report and the breakout appendices.

What Key Recommendations Came from the Workshop to Address the R&D Challenges?

Successful application of ML algorithms and big data approaches to the subsurface requires the involvement of subject matter experts (SMEs) and an outcomes-based approach. It is not uncommon for data scientists who have little to no geologic or reservoir engineering expertise to take broadly applicable datasets collected by the oil and gas industry and apply mature DA and ML tools to these datasets. Unfortunately, such exercises often end in failure or reflect only marginal success. When data scientists and SMEs who have experience in modeling and/or characterizing the subsurface collaborate (see Figure 1), the chances of success are much higher. Because subsurface problems are so complex and have high dimensionality in terms of the number of independent and dependent parameters, geoscience experts are needed, at the very least, to help shape and constrain the initial algorithms used.

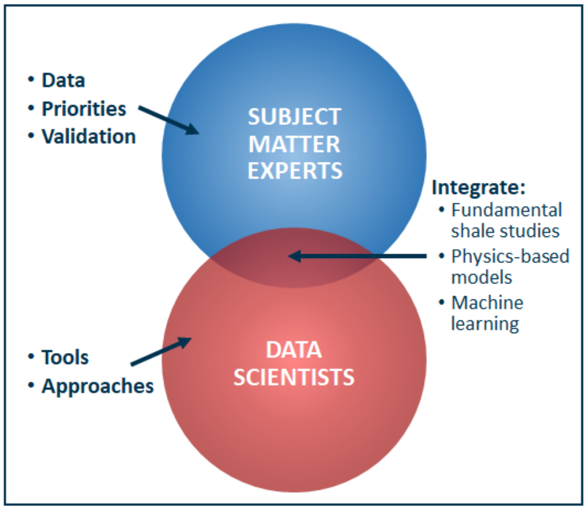

Figure 1. Geoscientists engage with data scientists to make DA and ML endeavors successful.

Additionally, engaging SMEs can help ensure that such work is outcome-driven—i.e., that the recommendations generated from the application of ML tools are focused on improved outcomes useful to operators. Having clear outcomes as drivers of the data collection and analysis process can help identify what ML approaches are needed to solve the problem (see Figure 2), and SMEs can help identify data sources that were not originally planned as part of the effort.

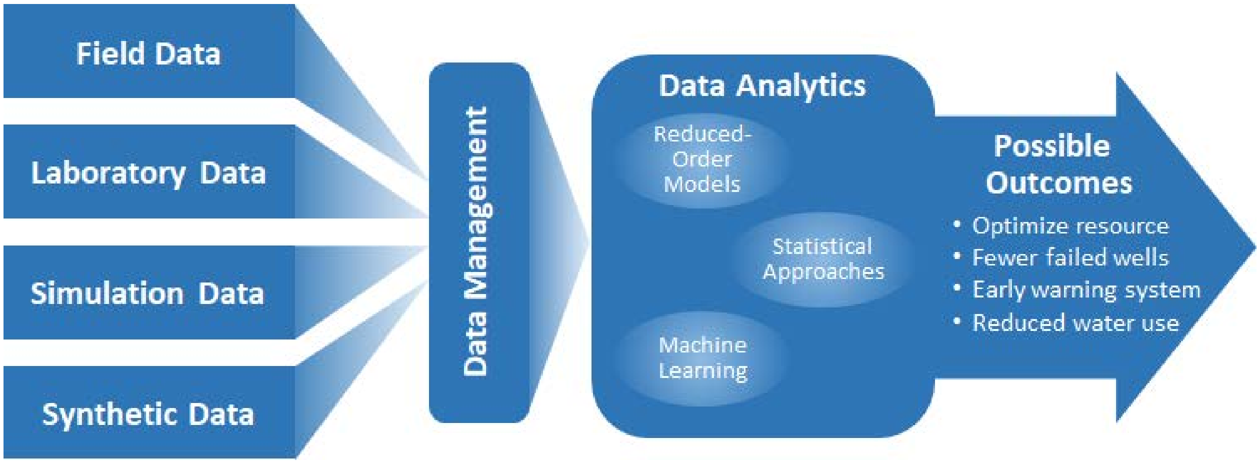

Figure 2. Multiple data sources from the field, laboratory, and simulations can be combined through strong DM tools to develop DA and ML tools that achieve targeted outcomes.

Field data should be supplemented with physics-based simulation to enable successful big data DA and ML approaches. It's often the case in subsurface applications that the amount and type of data collected are not comprehensive enough to do a complete analysis based on limited data streams alone. Analyses that rely on too little data, or on low-quality data, can easily produce misleading or erroneous results. Modeling may be needed to sample subsurface cases that are not sampled by well data alone. When the physical properties and chemistry of a subsurface reservoir can be modeled with high-fidelity, physics-based simulators, then the results from such simulations can be used to augment field data in ML applications (see Figure 2). In some cases, this is done by using the simulated data to help constrain the inputs, combination of inputs, or boundaries of the DA algorithms that are used. In cases where some of the data that have been collected is of questionable quality, reservoir simulation can help provide a quality control check on the data. Additionally, reservoir simulations can generate synthetic data that can be used to train and validate ML algorithms in similar ways regarding how field data are used. Some recent studies have shown that this approach is useful and sometimes necessary to get meaningful results from ML applications [1, 2]. Similarly, laboratory studies, including those focused on the physical properties and chemistry of a reservoir, could be used for training and validation procedures in the same way simulations are used. Further, laboratory studies can also be used to ground truth simulations and to test basic physics when it is unfeasible in situ.

Through the use of physics-based models and approaches, it will be possible to use ML approaches to significantly extrapolate from a mature and well-understood formation and into a new one within the same play or basin. By their very nature, DA approaches are only guaranteed to perform well when they are used for interpolating. However, with the combination of physics-based modeling and DA, it may be possible to use these approaches to infer key information within a given play area where limited exploration and data collection has occurred.

Access to large datasets, often owned by more than one operator, will be necessary to perform meaningful ML analyses. In the oil and gas industry, site characterization data are often held tightly and not shared with others because of the value that it is perceived to have. However, in many cases, a single operator might not have sufficient data assets in a single field or play to justify the use of ML and advanced DA. Recent studies have shown that, for some applications, data from several hundred to a thousand wells may be needed before DA can inform the operator in a meaningful way [3]. When it makes sense, data sharing or data pooling should be encouraged to create larger datasets that can provide significant new insights to field operators. Even when data pooling among operators is not necessary, researchers are going to require access to significant volumes of data from operators; for this to be successful, operators must be willing to share resources but should also expect to receive some type of ancillary benefit as a result of the research performed. Secure agreements that can allow researchers access to data while ensuring that proprietary data remain secure will be necessary as new techniques are applied and developed for specific use cases. The use of open-source software and platform agnostic approaches are likely to create the easiest and fastest environment for sharing such data and for the subsequent analysis, as long as those tools remain secure. Open-source software is also attractive because it can allow oil company clients to append reservoir-specific code.

New DM tools and approaches will be needed to enable broad application of DA and ML to subsurface applications. A wide variety of datasets need to be brought together (e.g., well logs, production logs, seismic surveys) to perform meaningful analyses. Unfortunately, data, even when of the same type, are often held in proprietary formats that are not easily accessible by others outside that organization. The challenge here is that the format for each of these types of data is wildly inconsistent by nature. Moreover, even similar data types (for instance, well logs) may vary in constitution, albeit subtle, from one organization to the next, and even within the same organization when collected at different times or by different groups. This predicament is a significant impediment to successful data mining and ML absent of diligent data preprocessing. If each data type had a standard format or even a few standard formats, that would make it much easier to meaningfully share data and information from one organization to another (e.g., operators, service companies, consultants, researchers). Standardization would also likely enable faster development of useful ML tools that can be more broadly tested and applied to industry problems.

In addition to formatting issues, there is also the concern about data volumes and transmission rates. With the rapid advances in sensing technologies that produce more and more data, it becomes much more challenging to transfer the high volumes of data collected at the wellhead onto the cloud to perform analyses. Several possible solutions to this problem have potential. For example, lossless data compression or similar techniques have a big impact when the techniques can reduce the massive datasets to just a few meaningful numbers. Edge computing (i.e., decentralized data processing) is another important tool to consider, as it brings the analytical power to the well, so that a significant portion of the analyses can be performed directly at the wellhead to extract only meaningful data for additional processing in the cloud. For example, an early warning signal can be identified at the wellhead from potentially several data parameters, and only a simple warning indicator can be passed back to the control room.

General Background, Scope, and Focus

The ability to monitor the subsurface for decision-making purposes is crucial to understanding and ultimately optimizing the exploration for and utilization of subsurface resources, as well as effectively monitoring operational safety, and diagnosing potential risks like wellbore integrity issues. This ability is critical not only for operators in the oil and gas industry and in carbon storage applications, but also for United States (U.S.) energy security and environment and public protection.

ML, DA, and DM have expanded rapidly in many commercial sectors, providing an array of resources that can be leveraged for subsurface applications. In addition, novel sensors, such as fiber optics-based sensors that measure distributed acoustics, temperature, and strain, are now able to provide signals in parts of the reservoir and at a temporal and spatial resolution not possible a few years ago. Nanosensors that can be injected into reservoirs, either as part of stimulation procedures or during drilling or injection operations, are also under development, providing the promise of an unprecedented look inside reservoirs at distances away from wells. Global positioning systems, interferometric synthetic aperture radar (InSAR), strain meters, continuous gravity measurement, continuous source seismic, and other approaches provide additional avenues for continuous and repeated sensing and imaging approaches that generate large volumes of data and can be broadly distributed across fields. Finally, developments in computational power and sophisticated physics-based simulation platforms have opened a pathway to physics-informed ML for complex subsurface systems, in turn enabling new strategies for the signature discovery and training of new algorithms that can provide the backbone for autonomous systems. In short, it is time to focus more stringently on the development of real-time decision-making capabilities that could transform the approaches commonly used for subsurface energy systems. Such transformational approaches may play an important role in supporting the goals of a broad range of stakeholders, including the DOE's Office of Fossil Energy. These goals include doubling resource recovery from unconventional systems, improving carbon storage capacity estimates, and ensuring the safe operation and long-term performance of all fossil energy-related subsurface activities.

However, while a high-resolution description of the subsurface enables better characterization of the phenomena and processes occurring in the subsurface, it also presents a challenge due to the large amounts of data that must be collected, transmitted, stored, and processed. Additionally, due to the complexity of these systems involving hydrologic, chemical, mechanical, and other processes, traditional physics-based computational models tend to struggle with the ability to fully capture all relevant physical/chemical phenomena in a timely manner. DA methods and ML approaches can support advancements in this area; the potential impact that this could have in subsurface monitoring and signal interpretation for real-time decision-making is significant.

The need to coordinate related R&D activities—to identify recent advancements, knowledge gaps, and beneficial future directions—has led to several recent activities by a variety of interested parties spanning industry, academia, and government. For instance, DOE's Office of Science held a workshop in February 2019 on scientific ML to define the challenges and opportunities for applied mathematics research and to increase the rigor, robustness, and reliability of ML for DOE mission requirements [4]. Early in 2018, the Center for Nonlinear Studies and the Center for Space and Earth Sciences at Los Alamos National Laboratory sponsored the Conference on ML in Solid Earth Geoscience, which focused on modern applications of ML in geoscience problems, including earthquakes, faulting, Earth imaging with a multitude of data types, geological characterization, and subsurface flow and transport. A second conference occurred in March 2019. In July 2018, the U.S. Energy Association hosted a workshop to identify how big data and ML can be leveraged to enable the advancement of coal energy systems, including carbon capture and storage [5]. This report applies to the "Real-Time Decision-Making for the Subsurface" workshop hosted by Carnegie Mellon University's Wilton E. Scott Institute for Energy Innovation in July 2018 and focuses on the R&D challenges inherent to big data approaches and analytics for subsurface energy applications.

Big data DA has the potential, through data mining and ML approaches, to help increase operational efficiencies by gathering actionable information and understanding hidden patterns and relationships in large, complex, multi-dimensional datasets. There are many challenges in this area, including the handling of large volumes of data, the velocity at which the data are generated and collected, and the variety of different data streams. Data-driven modeling and ML apply statistical techniques to construct predictive models that "learn" based on data streams from real-world activities without explicit modeling/programming of the underlying causal relationships. This is an important technique when sufficient domain knowledge (such as physics) is lacking to build a theoretical model capable of supporting computational simulations, or when a rapid predictive capability is needed. It also has many benefits, as it can often capture non-linear relationships between variables, avoid explicitly defining variable transformations/data models, automatically handle correlation between predictors, and support guided/automated tuning of model parameters.

ML is often successfully used when developing reduced-order models (ROMs). When a process or multicomponent system is particularly complex and computationally expensive to model (as is often the case for many subsurface phenomena), the system can be divided into discrete components, which can be validated separately against field and lab data. At this point, ML approaches can be applied to create a ROM that can rapidly reproduce the component model prediction. In subsurface applications, this method can be used to develop surrogate reservoir models, which are approximations of full-field models developed to accurately represent a given full-field reservoir simulation model. These models address many time-consuming operations performed with reservoir simulation models and have the potential to reduce the amount of time needed for reservoir simulations from days (via traditional physics-based, full-field model) to seconds (via ROM). Additionally, these models can be trained with a relatively small number of reservoir simulations or field/lab data, which is critical when time and resources are constrained. An example of a recently developed surrogate reservoir model can be found within the paper by Shahkarami A., Mohaghegh S. and Hajizadeh, Y. [6] In this work, a developed surrogate reservoir model using artificial neural networks (ANNs) was shown to have high accuracy in mimicking the behavior of a full-field reservoir simulation model.

Examples of successful application of ML techniques in oil and gas are rapidly growing, particularly in the areas of exploration and production, digital oil field management, predictive maintenance, and natural language processing. A few examples related to optimization of well placement include the study from Montgomery and O'Sullivan [7]. This study highlighted the importance of using the appropriate model in characterizing the impact of well location and spacing on well productivity gains—a factor that was highly underestimated in previous analyses. Similarly, Shahkarami and Wang [8] have coupled a data-driven predictive model with an economic model to evaluate influencing factors in net present value, such as well spacing and well interference, specifically for horizontal well spacing scenarios and hydraulic fracturing design.

A successful application of a purely DA approach with production well history matching for oil and gas production in the North Sea [6] has enabled faster prediction of reservoir behavior than traditional reservoir models. In theory, this approach, which relied on information from other wells in the formations and early production data, captures as much or more of the underlying physics as the physics-based models. In other instances, DA techniques have been used on specific processes in oil and gas production. An example includes a plunger lift optimization process where the addition of real-time acoustic monitoring was used to supplement available data, and a statistical DA approach was used to optimize timing of artificial lift operations to maximize recovery.

The detection of induced seismicity events has also benefitted from DA approaches. Skoumal et al. [9] developed a Repeating Signal Detector, a computationally efficient algorithm that uses agglomerative clustering to identify waveforms buried in years of seismic recordings using a single seismometer. Rouet-LeDuc et al. and Hulbert et al. [10, 11, 12, 13] showed that continuous seismic data processed by ML techniques can reveal information about a fault's physical properties, such as friction and/or displacement in laboratory and field studies. They also showed that the timing of upcoming failures can be bounded by the information contained in the signals [10]. These are a very limited number of the large group of examples that can be currently found within the open literature. Additionally, several projects are currently ongoing at many of the national laboratories, particularly projects on novel ML techniques for reservoirs production prediction, fractures development, and induced seismicity management.

As this field of study evolves, it is important to identify the goals specific to subsurface applications. Near-term goals should aim to incrementally (e.g., by 10%) improve resource recovery and storage potential and should include the reduction of interference in hydraulic fracturing stages, reduction of operational costs by removing, for example, costly low-performing stages, clusters and/or perforations, and the improvement of plume tracking capabilities. This will require the use and further development of current and novel sensing technologies as well as computational techniques and methods. Continuing to apply and find innovative ways to integrate existing DA approaches, as well as current computational capabilities, will also be needed. This is the first step needed to achieve the ultimate, more ambitious goals of the applications of DA to subsurface, which could include doubling oil and gas recovery from unconventional reservoirs, doubling carbon storage resource potential, and increasing EOR potential by 25% within the next 10 years. These transformational approaches will require the incorporation of novel continuous and distributed (e.g., distributed acoustic sensing [DAS], nano) sensing and the leveraging of high-performance computing technology (e.g., exascale).

The achievement of these goals relies on the ability to overcome a series of challenges, including the lack of access to data, especially labeled data; the need for better approaches to combine high-performance models, ROMs, DA, and ML; and the need for improved visualization of large multi-dimensional subsurface datasets. Additionally, challenges in data fusion will need to be resolved by combining multiple types of datasets into a single analysis (discrete, continuous, point, distributed), and by integrating continuous monitoring measurements with periodic measurements and reservoirs simulations.

The rapid integration of big data DA and ML in subsurface applications requires a coherent strategy and a clear definition of the questions surrounding the field with new, information-rich datasets. Part of the effort should go into understanding how researchers can use data and complex models to answer these questions.

The remainder of the main body of this report focuses of information and themes that were common among the three sub groups included in the workshop, specifically

1. Tools and approaches available for DM and analysis;
2. Methods for selecting and applying such analytical tools; and
3. Data needs, and challenges to be overcome.

A great deal of information was covered during the workshop within the three breakout groups (i.e., resource recovery and utilization, autonomous monitoring, and stress and seismicity), but because of the significant variability regarding the applications, data needs, and challenges among these three groups, their summaries are presented separately in appendices 1–4. Appendix 5 summarizes the attendees at the workshop; Appendix 6 provides the workshop schedule.

Enabling Tools and Approaches

In the oil and gas industry, hydrocarbon production from unconventional formations has skyrocketed in the past decade. Despite the prolific production, the vast majority of oil and gas remain in the ground. For both unconventional oil and gas exploration and production and carbon storage, the risk or perception of risk can be the biggest barrier to success. Additionally, a significant impediment to full development in both areas is the cost of operating and managing these fields. The focus in unconventional reservoirs has customarily been to produce large volumes as quickly as possible, ensuring a short payback period for capital expenditure. In return, the rapid subsequent decline in production in outbound years has resulted in these projects being overall less profitable than comparable deep-water conventional reservoirs. Often, the unconventional wells are shut-in or abandoned after only a few years of production. A considerable challenge for DOE is to develop and deploy technology to increase the longer-term production of such wells.

Given the recent and continuing advances in sensing and computational capabilities, there will be opportunities for transformational changes in oil and gas production and carbon storage activities through the use of DA and ML. Such technologies have the potential to enable a 100% or greater increase in oil production from unconventional formations, to detect safety or containment issues before they occur, and to significantly improve efficiencies in operations and reduce their overall costs.

In the previous section, several use cases in both unconventional oil and gas and carbon storage were discussed. More specific examples are outlined in further detail in the following subsections, as well as in the appendices of this document. In the following subsections, several of the potential tools and techniques that can be used to meet these challenges are addressed with examples from specific use cases from subsurface energy system applications.

Tools and Techniques

Analysis of subsurface systems using data-driven modeling (like ML) provides capability, tools, and approaches that facilitate new developments pertaining to: (1) acquiring and managing data in large volumes, of different varieties, and at high velocities; and (2) the use of statistical techniques to “mine” the data and discover hidden patterns of association and relationships in large, complex, multivariate datasets [14, 15, 16]. The ultimate goal is to develop data-driven insights for understanding and optimizing the performance of engineered subsurface systems [16]. The development of such data-driven models typically follows a process of distinct and iterative steps that consists of problem classification through data collection, algorithm selection, validation, and execution. At the data acquisition and processing stage, there are possibilities for integrating a multitude of data sources typically common to disparate components associated with understanding the behavior and performance of engineered natural systems. For instance, geologic characterization data (i.e., cores, logs, and seismic survey data), fluid/geochemical sampling data, monitoring-based data, as well as production (i.e., water, hydrocarbons) and injection data are all relevant types of data being generated at both commercial and pilot field sites (across both the oil and gas spectrum and CO_2 storage field projects) that could inform data-driven model development. A typical process outline for a ML application is presented in Figure 3, which highlights prominent stages [17, 18, 19]. However, despite the relatively sequential setup of Figure 3, the process in general is likely to be highly iterative.

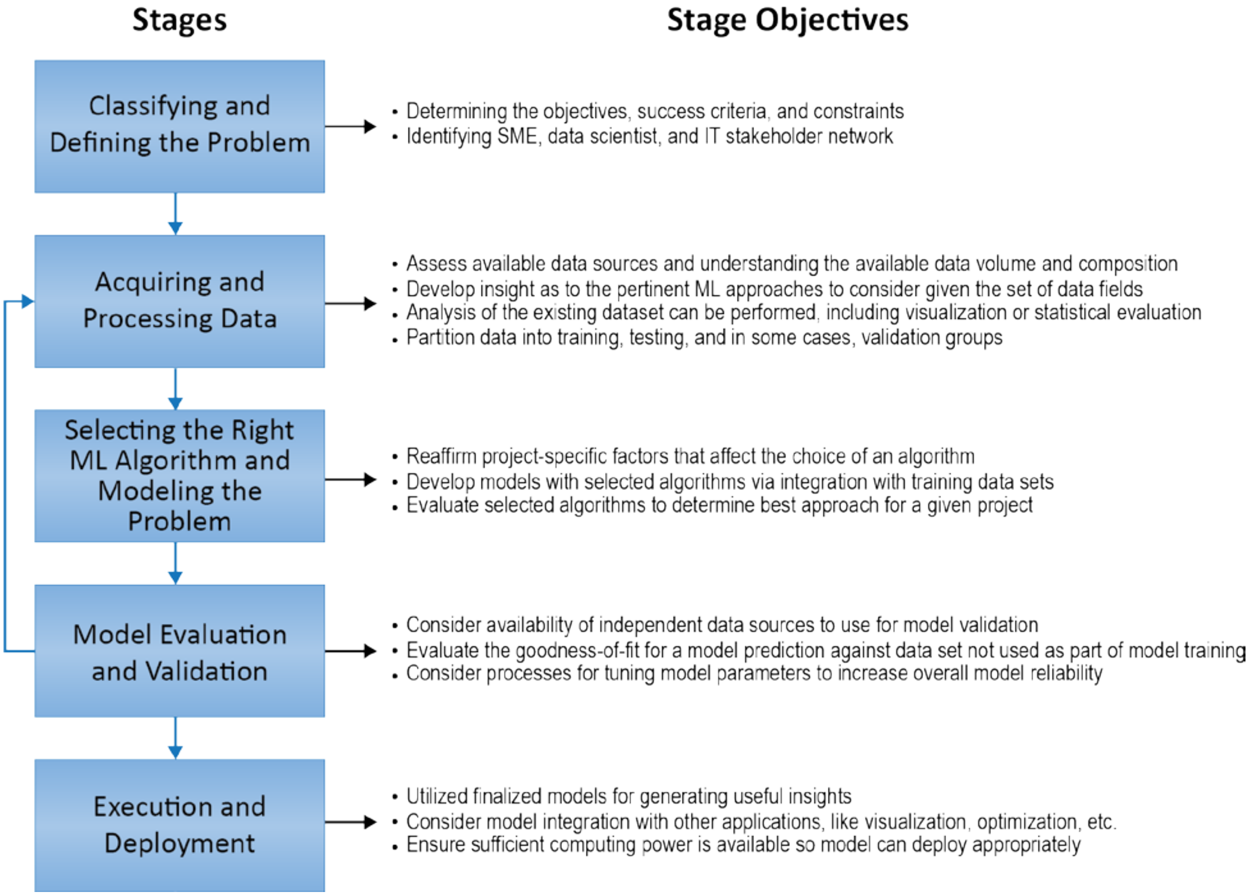
Stages
Stage Objectives
Classifying and Defining the Problem
Determining the objectives, success criteria, and constraints
Identifying SME, data scientist, and IT stakeholder network
Acquiring and Processing Data
Assess available data sources and understanding the available data volume and composition
Develop insight as to the pertinent ML approaches to consider given the set of data fields
Analysis of the existing dataset can be performed, including visualization or statistical evaluation
Partition data into training, testing, and in some cases, validation groups
Selecting the Right ML Algorithm and Modeling the Problem
Reaffirm project-specific factors that affect the choice of an algorithm
Develop models with selected algorithms via integration with training data sets
Evaluate selected algorithms to determine best approach for a given project
Model Evaluation and Validation
Consider availability of independent data sources to use for model validation
Evaluate the goodness-of-fit for a model prediction against data set not used as part of model training
Consider processes for tuning model parameters to increase overall model reliability
Execution and Deployment
Utilized finalized models for generating useful insights
Consider model integration with other applications, like visualization, optimization, etc.
Ensure sufficient computing power is available so model can deploy appropriately

Figure 3. Schematic depicting stages pertaining to the development of data-driven ML models [17].

Since the development and coding of advanced statistical algorithms, there has been an increased reliance on the use of commercial packages (such as SAS) and open-source packages (R and Python). As a result, ML algorithms are readily available to the larger community. However, there remains the need to implement sound approaches for integrating these types of algorithms as part of any application, not just specifically for subsurface-related applications. Mishra and Datta-Gupta [14] have identified the following items as critical considerations in this regard: (1) choosing the right algorithm(s) for the problem as opposed to using a preferred one for all cases, (2) applying the algorithm(s) with the proper choice of user-defined parameters, (3) avoiding the problem of data over-fitting and resulting bias in fitted model predictions, and

(4) ensuring that the data-driven model makes physical sense in terms of variable selection and parameter importance. In the following subsections, commonly used data-driven modeling tools and techniques applicable for petroleum geosciences and subsurface applications are discussed. As indicated in Figure 3, a typical process of creating a model that applies the best tools and techniques to effectively evaluate the specific problem case consists of several stages.

Classifying and Defining the Problem

The first stage is classifying and defining the specific problem, which involves determining the objectives, success criteria, and constraints used in the modeling process. Building and operating an end-to-end ML-based system also requires stakeholders comprising SMEs, data scientists, and possibly information technology operations personnel [17]—all of which provide unique perspectives to effectively evaluate and address the problem of interest [20]. Depending on the problem evaluated, the mix of stakeholders engaged, and the approaches considered could vary significantly.

Several recent studies have successfully demonstrated the use of ML and DA in subsurface applications. These studies have explored topics pertaining to predicting hydrocarbon production in unconventional reservoirs [21, 22, 23], lithofacies identification and characterization [24, 25, 26], CO_2 storage monitoring design [2], and CO_2 storage monitoring anomaly detection [27], to name a few. However, other possible uses for data-driven models in subsurface-related applications could include risk assessment, informing real-time decisions in the field, understanding seismicity and dynamic stress state in the subsurface, well design and completion optimization, informing drilling and geo-steering, and tracking CO_2 and pressure plumes in the subsurface [20]. Ultimately, the problem being addressed will directly influence the types (and possibly the quantity) of the data needed, the algorithm(s) selected, the types of SMEs best suited to inform decisions, and the approaches toward validation.

Acquiring and Processing Data

Given its inherent data-driven nature, the quality of any ML effort is based on the quality of the data it uses. The quality and quantity of data gathered will directly impact how well the predictive model developed will ultimately perform [28]. Once the specific problem of interest has been defined and objectives of the effort established, the next step would involve assessing available data sources and understanding the available data volume and composition. Depending on the specific application, project stakeholders can begin to gain insight regarding the pertinent ML approaches, given the set of data fields and parameters and volume of historical data available or that could become available [18]. For subsurface applications, possible data types for a given project may include some combination of the following: (1) data from field projects, (2) exploratory and laboratory-generated data, (3) simulation-derived data (from physics-based or ROMs), and (4) synthetic datasets [20]. One key difference in data produced by ML applications when compared to more conventional data analysis is that its most valuable data are typically raw data absent from excessive aggregation [18]. However, there are noted approaches for normalizing, imputing, transforming, and dealing with outliers in datasets when needed to make datasets more palatable for ML [29, 30]. Subsurface datasets are expected to be large and possibly continuous in nature (as field projects progress and models and laboratory experiments generate new data). To provide an example for data types relevant in subsurface applications, the parameter sets from studies performed by Shih et al., [22] and Schuetter et al. [21] are highlighted in Table 1 below. These studies are similar in that the authors utilized ML algorithms to build predictive models during the first 12 months of cumulative production (gas in western Marcellus Shale in Shih et al., and oil in Delaware Basin in Scheutter et al.) in unconventional reservoirs based on well completion datasets and geologic properties in the associated study areas. The datasets from these two studies share several commonalities in the parameters of interest; however, there are also unique fields associated with each study (which could be from the variability of available data across study regions).

Parameter Type	Parameter	In Shih et al. 2018	In Scheutter et al. 2015
Response	First 12 months cumulative production	X	X
Technology	Perforated lateral length	X	X
	Total water used for hydraulic fracturing	X	X
	Proppant amount	X	X
	Additive used for hydraulic fracturing	X	X
	Well azimuth trajectory	X	X
	Well spacing	X	
	Pad drilled (Y/N)	X	
	Well completion year		X
	Stages		X
	Drift angle		X
	Proppant concentration		X
	Non-disclosed operator parameter		X
Geology	True vertical depth	X	X
	Location data (latitude and longitude)	X	X
	Thickness	X	
	Gas to oil ratio cumulative at 12 months	X	
	Gamma ray (GR)	X	
	Thermal maturity	X	

Table 1. Comparison of data parameters used in two subsurface ML application studies.

It is important to note that both studies were built around achieving similar objectives but utilized different configurations of data. In the Shih et al. study specifically, the data spanning production, well design, and reservoir geology were merged from disparate sources to form the finalized dataset [22]. Schuetter et al. additionally included several operator-specific data parameters used as predictor variables likely not available from public datasets. However, those parameters were masked and given non-descriptive identifiers. In subsurface applications, particularly related to oil and gas, a majority of potentially useful and viable data is often not publicly available. Collaboration with industry is one way to potentially expand the availability of existing datasets moving forward.

Once data are gathered, it typically requires some level of preparation and processing for efficient integration into ML algorithms. Additionally, analysis of the existing dataset can be performed, including visualization or statistical evaluation to help gain insight into relevant relationships between the different variables, as well as determining if there are any data imbalances present [22, 21, 28]. These steps typically comprise a substantial portion of the overall effort in an ML application and necessitates considerable human intervention as part of the process.

Finalized datasets can then be partitioned into training, testing, and in some cases, validation (or calibration) groups [22]. The training group is intended to train the model using a given ML algorithm and develop a preliminary “trained” predictive model. The test group is used to evaluate the predictive accuracy performance of the trained model against unseen examples of data for which the model was trained. The validation dataset group provides another set of unseen example data separate from the training set that can support the tuning and optimization of algorithm hyperparameters and other components of a given algorithm(s) architecture. This dataset can also be used to validate models to avoid over-fitting or under-fitting [22, 28] (discussed in Model Evaluation and Validation). The validation step typically occurs after initial model training, and prior to model testing. Approaches to data splitting tend to vary across published studies and may likely vary depending on the application. However, two general rules apply to this process: (1) ensure dataset groups are large enough to yield statistically meaningful results, and (2) ensure each grouping be representative of the dataset as a whole [31]. The Shih et al. [22] and Schuetter et al. [21] studies each provide additional insight into approaches to data splitting for subsurface-related ML applications. For the approach utilized by Shih et al. (Figure 4), the finalized dataset, containing 1,418 wells, was split into training, validation, and testing groups in a 60/20/20 percentage split that was randomly determined. Schuetter et al. utilized a “k-fold” cross-validation approach, schematically depicted in Figure 5 [21]. In this approach, the training group is randomly split into different “k” groups or “folds.” Next, each of the “k” groups are held out one at a time, and the model is trained on the remaining “k-1” groups. The trained model is then used to make predictions on the group that was omitted. After cycling through all “k” groups, there will be a single cross-validated prediction for every observation in the dataset, in which the predictions are made using a model whose observation was not included in the training set [14].

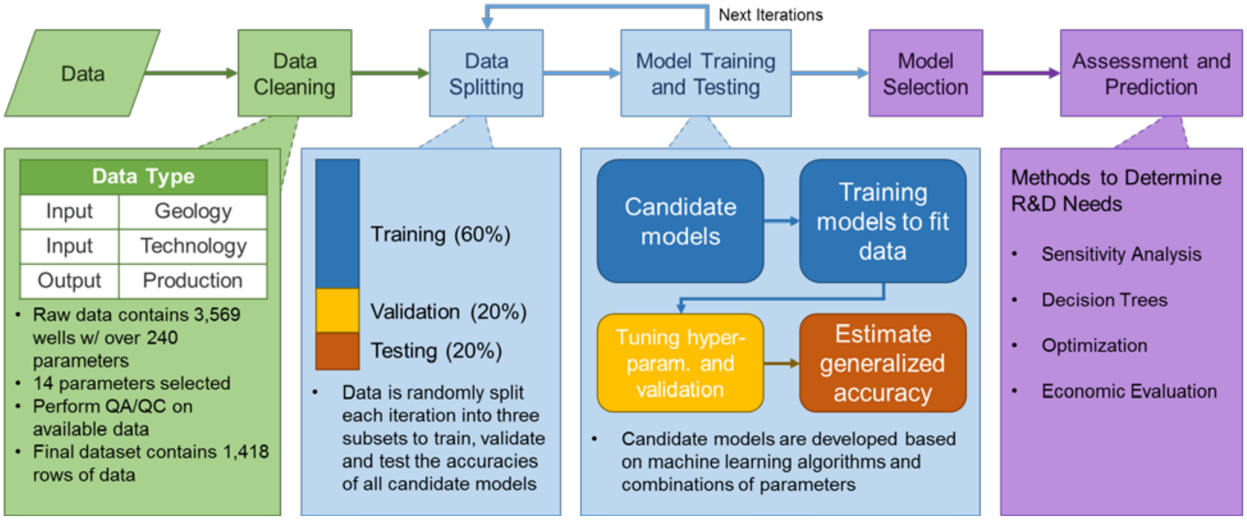
Next Iterations
Data
Data Cleaning
Data Splitting
Model Training and Testing
Model Selection
Assessment and Prediction
Data Type
Input | Geology
Input | Technology
Output | Production
• Raw data contains 3,569 wells w/ over 240 parameters
• 14 parameters selected
• Perform QA/QC on available data
• Final dataset contains 1,418 rows of data
Training (60%)
Validation (20%)
Testing (20%)
• Data is randomly split each iteration into three subsets to train, validate and test the accuracies of all candidate models
Candidate models
Training models to fit data
Tuning hyper-param. and validation
Estimate generalized accuracy
• Candidate models are developed based on machine learning algorithms and combinations of parameters
Methods to Determine R&D Needs
• Sensitivity Analysis
• Decision Trees
• Optimization
• Economic Evaluation

Figure 4. Regression modeling and sensitivity analysis framework used as part of the Shih et al. study supported by the National Energy Technology Laboratory [22].

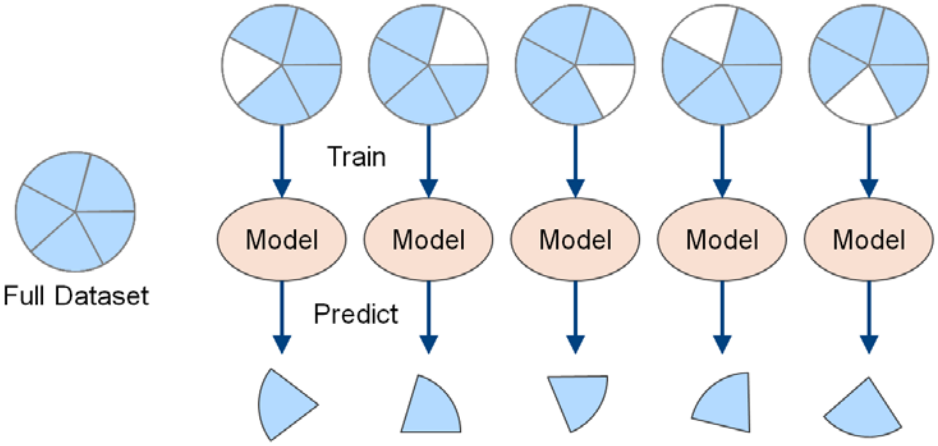

Figure 5. Conceptual representation of k-fold the cross-validation with k=5 [21, 14].

It is important to note that this type of cross-validation procedure can be extended by repeating the entire process with a different random selection of training and testing/validation groups, as performed by Schuetter et al. [21]. The result of repeated cross-validation using different repeated runs of randomly selected groups will yield several unique predictions on each of the observations. These can not only be aggregated to compute statistics on quality-of-fit metrics, but they also give important insight into the variability in model predictions depending on the characteristics of the training set [14].

The examples here are not fully inclusive of all the possible approaches to processing data but provide a perspective of examples implemented in subsurface applications. The next step involves choosing the range of ML algorithms that best suits the objective of the project.

Selecting the Right ML Algorithm and Modeling the Problem

Given the ML algorithms available, there is typically no standalone solution or singular approach that fits all applications. This is especially true in the subsurface arena, where there is growing interest in using DA to improve upon the currently established best practices. There could be several factors that can affect decisions regarding algorithm selection, which are dependent on the nature of the problem investigated. However, once the objectives of the project have been clearly defined and a strong understanding of the data has been gained, decisions toward algorithm selection should be more straightforward. Some of the factors affecting the choice of an algorithm are [30]

- Alignment to project objectives and goals
- Volume of pre-processing required
- Prediction accuracy
- Model speed, both in time needed to construct and to make predictions
- Model scalability

ML typically falls under three prominent categories: supervised learning, unsupervised learning, and reinforcement learning. Prominent ML algorithms and approaches available are aligned specifically to these categories [32]:

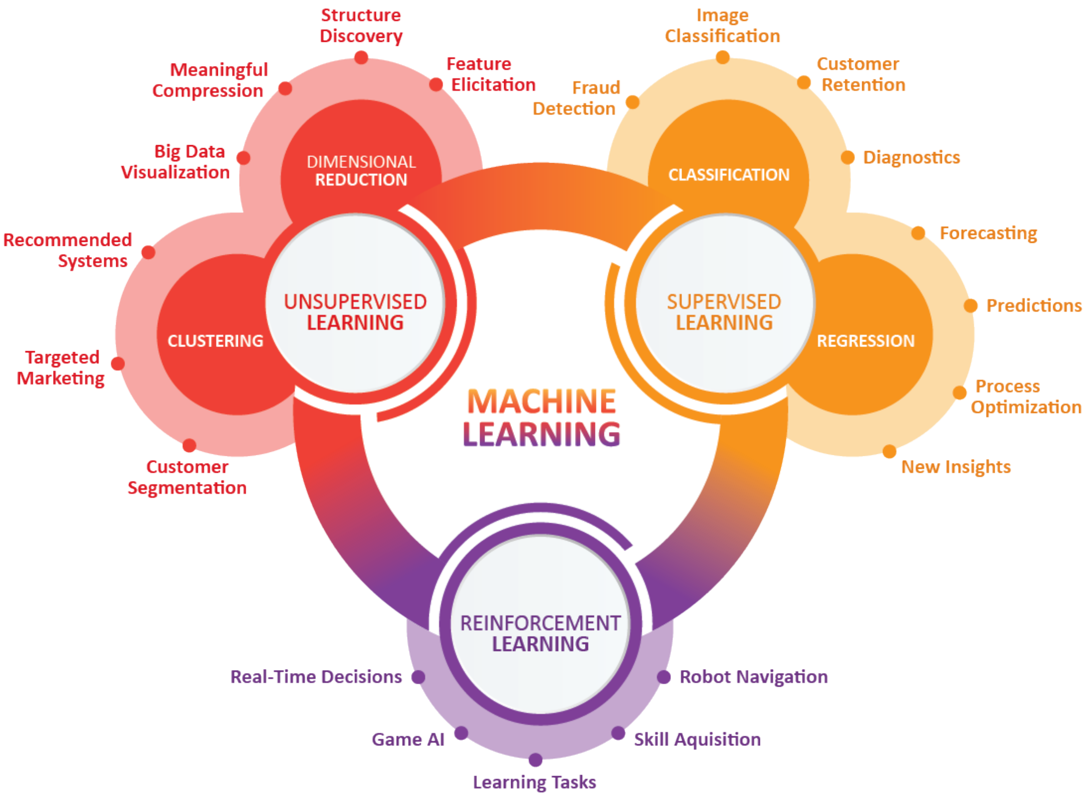

Figure 6. Schematic of the ML concept with associated categories and example applications [33].

Supervised learning utilizes training datasets comprising input-output labeled data pairs that are analyzed through ML tasks to predict an expected outcome. Data labels need to be assigned under human supervision, giving this category its namesake. Supervised learning can be further subdivided into (1) regression problems, where the response variable is continuous; or (2) classification problems, where the response variable is categorical. For both cases, predictor variables may be continuous and/or categorical [14].

As mentioned in the bullets above, the appropriate algorithm(s) ultimately selected will be based on several factors. Figure 12 in Appendix 4 provides further insight into selection of the most appropriate algorithm based on problem definition and data available/expected.

Selected candidate algorithms can be utilized for the creation of a trained model that will best translate to new input and response data. Essentially, training datasets are assigned to an algorithm in order to develop a model capable of making reliable predictions on both new or untrained (i.e., testing datasets) data. Once models have been developed from training, model evaluation and validation steps are needed to assess overall performance and effectiveness (discussed in more detail in Model Evaluation and Validation) [17].

ML algorithms that hold the most application potential can be selected for modeling training on project data, run either in parallel or series, and at the end evaluate the performance of the algorithms to select the best one(s) [30]. Bhattacharya, et al. [24] implemented such an approach by using four ML algorithms to develop models that would predict specific lithofacies from well logs in the Bakken and Mahantango-Marcellus Shale formations. A similar approach was performed by Shih et al. [22], in which nine different ML algorithms were tested to gauge performance in predicting gas production in a selected area in the Marcellus Shale (Figure 7).

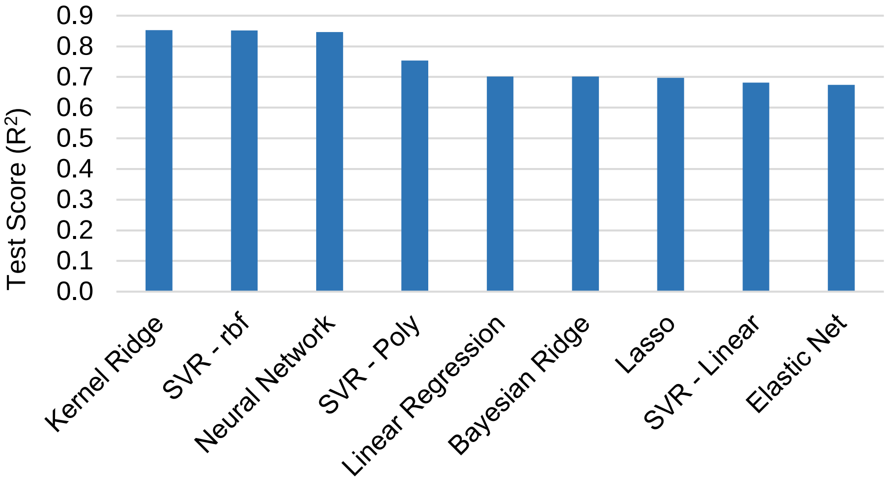

Figure 7. Evaluation of model performance trained using different algorithms (using R^2 for actual vs. predicted values) against testing data as part of the Shih et al. study [22].

Overall, this stage of a ML project is inherently intertwined with model validation and evaluation, ultimately because the selected model must perform effectively. The specific ML application and associated data, objectives, and priorities will ultimately help inform the selection of an appropriate algorithm, but evaluation and validation are needed to ensure model efficacy.

Model Evaluation and Validation

Once training is complete on selected algorithms, developed models can be evaluated for overall utility (specific approaches and examples presented in Appendix 4). Evaluation and validation allow the testing of the model against representative datasets of real-world phenomena the model is attempting to mimic or predict [28]. For model validation, a common approach pertains to quantifying the quality-of-fit for a model prediction against datasets not used as part of model training, such as the testing dataset or a separate independent test set (possibly derived from a physics-derived model).

If the developed model underperforms, it might be necessary to utilize more advanced strategies to improve the performance of the model, utilize a different algorithm or algorithms, and possibly supplement datasets with additional data (either in quantity or complexity). Further improvement can possibly be accomplished by tuning certain model parameters. Examples of such tuning parameters include: (1) number of variables randomly sampled as candidates at each split and number of trees for the random forest algorithm, (2) number of trees for the gradient boosting regression algorithm, (3) cost parameter for the support vector machine algorithm, and (4) number of hidden layers and hidden units for the ANN algorithm [16, 14] (discussed in more detail in Appendix 4).

Fitting data-driven models with large, multivariate datasets often results in complex interactions of variables with one another that are often nonlinear. As a result, there is an inherent challenge in developing a straightforward understanding of input-output relationships and key sensitivities based on a simple evaluation of model results [14]. As part of this validation and evaluation phase, one can begin to investigate the importance of variables on the predicted responses to target key factors for further investigation. There are several possible approaches available for investigating parameter importance [16, 14], but a straightforward approach for variable importance that is not tied to any particular model is based on the concept of R^2-loss. This method works for regression-based models and is centered on removing a predictor parameter from the training dataset and noting the accuracy loss of that model [14]. The accuracy loss is a change in the R^2 value between actual and predicted values from the finalized model and the test model when one data parameter is removed. Parameters of greater influence, in theory, will have a larger noted effect on accuracy loss. Alternatively, if a superfluous predictor is removed from the model, there should be little to no impact on the accuracy (Figure 8).

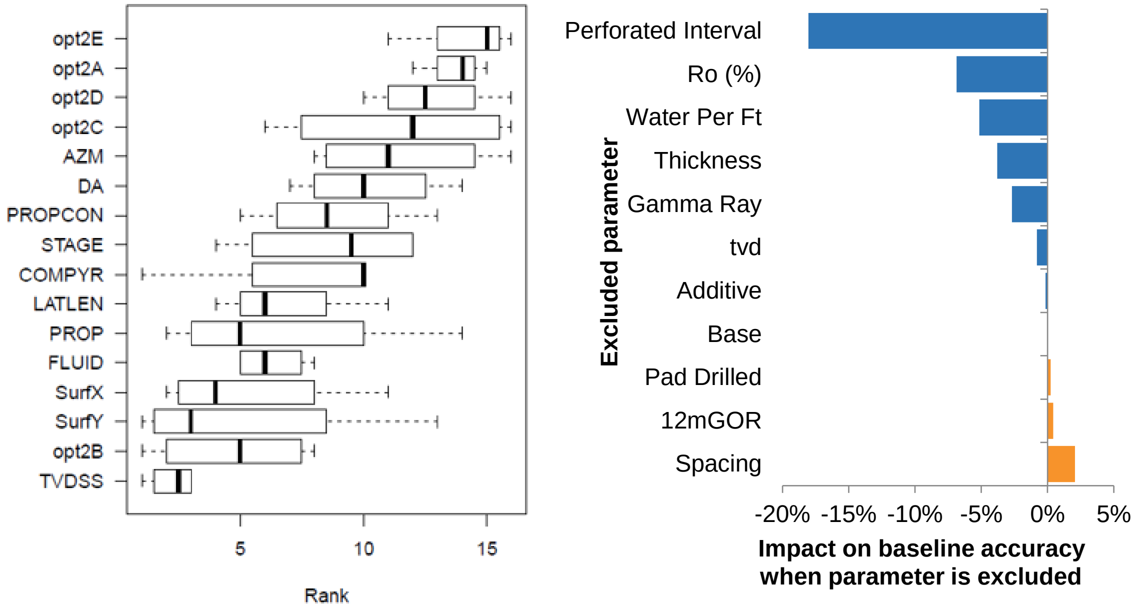

Figure 8. Example R^2-loss parameter impact evaluation from Scheutter et al. [21] (left) Shih et al. [22] (right).

Model parametric impact interpretability is best explained when model response changes as the value of all of the predictor variables are changed/omitted. This is similar to a standard one-parameter-at-a-time (OPAT) sensitivity analysis [14]. Other strategies include performing conditional sensitivity analysis to quantify the model response for a specified variation of any given predictor, when other possibly correlated inputs are varied, and the uncorrelated input parameters are fixed at some reference values [14]. This approach helps to determine collinearity among data parameters, which presents an advantage over OPAT analysis. For example, both the Shih et al. and Scheutter et al. studies identified strong collinearity between proppant and fluid used in well completions [22, 21]. In clustering-based approaches, classification trees and partitioning plots are not necessarily explicit parametric impact quantification approaches, but are example approaches which depict parameters and associated threshold values that separate pre-defined groups (i.e., high- vs. low-producing wells).

Execution and Deployment

Models that perform satisfactorily can be deployed for their intended task and ultimately help generate useful insights. ML technologies offer the benefits of speed, power, and efficiency through learning from data and trends without having to explicitly program these characteristics into an application [17]. Developing and deploying ML technology in subsurface applications has the potential to provide more efficient and accurate analytical methods that can ultimately transform petroleum geoscience into a much more data-driven science and enable better understanding of engineered subsurface system performance.

Understanding and utilizing the information gained from such approaches, tools, and/or new functionality is an important next step. For instance, deployed outputs could take the form of reported insights, supplemental data for existing physics-based models, or information to be stored or fed into other systems. Regarding the latter point, ML applications have enormous potential to integrate with other capabilities, including visualization (critical for enabling "human-in-the-loop"), optimization of modeled systems [22], and potentially autonomous monitoring systems [20].

One last key consideration relates to the amount of processing power that will be needed to effectively execute ML routines. Depending on the specific ML routine, the computational performance needed could be significant. For example, a simple neural net with only four or five inputs would be expected to handle comfortably using a regular central processing unit on a desktop server or laptop computer. However, a net that has numerous features—designed to perform advanced, "deep learning" routines—will likely need high-throughput computing power on the execution platform in the form of high-performance computing clusters, or compute kernels executing on high-powered graphics processing units [17].

Strategies to Address Barriers and Missing Data

Even with the best standard DA and ML practices, it is often the case that there is insufficient data to capture all of the physical, mechanical, and chemical processes that occur in the field. Several barriers were identified that may need to be overcome before transformational goals can be achieved. These barriers range from organizational and logistical issues to significant technical challenges, including lack of data on key processes or the inability to share data; insufficient bandwidth for transmitting data; heterogeneity of hardware, software, and data formats; and insufficient labor resources with regard to subject matter expertise and data science experience. Several different strategies can be used to overcome these potential barriers.

One of the key strategies that was identified to help with missing data or incomplete datasets was the incorporation of physics-based models into the process. In some cases, disparate outcomes seem to relate to the same inputs, and physics-based models can be used to help better sort data that comes from the subsurface into different categories, as well as better constrain the ML algorithms. In addition, physics-based models can be used to directly train ML algorithms to generate rapid predictive capabilities. Incorporation of physics-based models can bring its own challenges, such as having the computational resources available to generate the training datasets with such high-fidelity models. High performance computing capabilities, where available, can help alleviate this challenge.

In addition to physics-based models, several other approaches can be used to help address these areas. Data fusion is needed to integrate multiple datasets, but this can be challenging when the data are of different types (e.g., temporal, spatial) and come in different formats. Developing data standards can go a long way to addressing some of these problems, but to date, such efforts have been unsuccessful. Edge computing and data compression can also be powerful ways of addressing some of the bandwidth issues inherent to transmitting large volumes of data. Data compression can reduce the volume of data that is needed, and edge computing can reduce the need to transmit data by performing the necessary computations in or near the wells themselves. This can introduce its own challenges associated with getting the appropriate algorithms and updates from a central location to the edge.

Data organization and management can be improved with numerous approaches. Although it may not be necessary, using an open-source and platform-agnostic approach ensures that the final products will be available to a wide range of users. Data collectors should be encouraged to generate better descriptions of the ontology of their data. Additionally, generators of platforms and algorithms should encourage findable, accessible, interoperable, and reusable (FAIR) principles.

References

[1] Bacon, D. H., Locke II, R. A., Keating, E., Carroll, S., Iranmanesh, A., Mansoor, K., Wimmer, B., Zheng, L., Shao, H., Greenberg, S., "Application of the Aquifer Impact Model to Support Decisions at a CO2 Sequestration Site," *Greenhouse Gases: Science and Technology,* vol. 7, pp. 1020-1034, 2018.

[2] Chen, B., Harp, D., Lin, Y., Keating, E., and Pawar, R., "Geologic CO2 sequestration monitoring design: A machine learning and uncertainty quantification based approach," *Applied Energy,* vol. 225, pp. 332-345, 2018.

[3] Jacobs, T., "Pioneer's analytics project reveals the good and bad of machine learning," *JPT Digital Editor,* vol. 70, no. 9, 26 July 2018.

[4] Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., and Willcox, K., "Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence," United States Department of Energy: Office of Science, Washington, D.C., 2019.

[5] United States Energy Association, "Big Data and Machine Learning for Clean Coal and Carbon Management Strategic Initiatives," United States Energy Association, 12 July 2018. [Online]. Available: https://www.usea.org/event/big-data-and-machine-learning-clean-coal-and-carbon-management-strategic-initiatives. [Accessed 26 March 2019].

[6] Shahkarami A., Mohaghegh S., Hajizadeh, Y., "Assisted History Matching Using Pattern Recognition Technology," *SPE-173405,* 2015.

[7] Montgomery, J., and O'Sullivan, F., "Spatial variability of tight oil well productivity and the impact of technology," *Applied Energy,* vol. 195, pp. 334-355, 2017.

[8] Shahkarami, A., Wang, G., "Horizontal Well Spacing and Hydraulic Fracturing Design Optimization: A Case Study on Utica-Point Pleasant Shale Play," *Journal of Sustainable Energy Engineering,* vol. 5, no. 2, pp. 148-162(15), 2017.

[9] Skoumal R., Brudzinski M., Currie B., "An efficient repeating signal detector to investigate earthquake swarms," *Journal of Geophysical Research,* vol. 121, pp. 5880-5897, 2016.

[10] Rouet-LeDuc, B., C. Hulbert, D. C. Bolton, C. X. Ren, J. Riviere, C. Marone, R. A. Guyer, P. A. Johnson, "Estimating Fault Friction From Seismic Signals," *Geophysical Research Letters,* vol. 45, pp. 1321-1329, 2018.

[11] Hulbert, C., Rouet-LeDuc, B., C. X. Ren, J. Riviere, D. C. Bolton, C. Marone, P. A. Johnson, "Estimating the Physical State of a Laboratory Slow Slipping Fault from Seismic Signals," *Cornell University. Physics, Geophysics,* 2018.

[12] Rouet-LeDuc, B., C. Hulbert, P. A. Johnson, "Breaking Cascadia's Silence: Machine Learning Reveals the Constant Chatter of the Megathrust," *Los Alamos National Laboratory,* 2018.

[13] Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C., and Johnson, P., "Machine Learning Predicts Laboratory Earthquakes," *Geophysical Research Letters,* vol. 44, pp. 9276-9282, 2017.

[14] Mishra, S., and Datta-Gupta, A., Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences, Amsterdam, Netherlands: Elsevier, 2018.

[15] Holdaway, K., Harnessing Oil and Gas Big Data with Analytics, Wiley, 2014.

[16] Mishra, S., and Lin, L., "Application of Data Analytics for Production Optimization in Unconventional Reservoirs: A Critical Review," in *Unconventional Resources Technology Conference*, Austin, Texas, 2017.

[17] Sapp, C., "Preparing and Architecting for Machine Learning," Gartner, 17 January 2017. [Online]. Available: https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/preparing_and_architecting_for_machine_learning.pdf. [Accessed 10 September 2018].

[18] Khaytin, A., "The five stages of machine learning implementation," EENews Europe, 9 January 2017. [Online]. Available: http://www.eenewseurope.com/news/five-stages-machine-learning-implementation. [Accessed 11 September 2018].

[19] Analytics Vidhya, "Artificial Intelligence Demystified," 23 December 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/12/artificial-intelligence-demystified/. [Accessed 10 September 2018].

[20] National Energy Technology Laboratory, "Data Analytics and Machine Learning Panel," in *Mastering the Subsurface Through Technology Innovation, Partnerships, and Collaboration: Carbon Storage and Oil and Natural Gas Technologies Review Meeting*, Pittsburgh, Pennsylvania, 2018.

[21] Schuetter, J., Mishra, S., Zhong, M., and LaFollette, R., "Data Analytics for Production Optimization in Unconventional Reservoirs," in *Unconventional Resources Technology Conference*, San Antonio, Texas, 2015.

[22] Shih, C., Vikara, D., Venkatesh, A., Wendt, A., Lin, S., and Remson, D., "Evaluation of Shale Gas Production Drivers by Predictive Modeling on Well Completion, Production, and Geologic Data," U.S. Department of Energy and the National Energy Technology Laboratory, Pittsburgh, Pennsylvania, 2018.

[23] Izadi, G., Zhong, M., and LaFollette, R., "Application of Multivariate Analysis and Geographic Information Systems Pattern-Recognition Analysis to Production Results in the Bakken Light Tight Oil Play," in *Society of Petroleum Engineers Hydraulic Fracturing Technology Conference*, The Woodlands, Texas, 2015.

[24] Bhattacharya, S., Carr, T., and Pal, M., "Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA," *Journal of Natural Gas Science and Engineering,* vol. 33, pp. 1119-1133, 2016.

[25] Zhao, T., Jayaram, V., Marfurt, and Zhao, H., "Lithofacies Classification in Barnett Shale Using Proximal Support Vector Machine," *Society of Exploration Geophysicists,* pp. 1491-1495, 2014.

[26] Zhao, T., Verma, S., Devegowda, D., and Jayaram, J., "TOC Estimation in the Barnett Shale From Triple Combo Logs and Time Series Analysis," *Society of Exploration Geophysicists,* pp. 791-795, 2015.

[27] Sun, A., "Development of a Framework for Data Integration, Assimilation, and Learning for Geological Carbon Sequestration (DIAL-GCS)," in *Mastering the Subsurface Through Technology Innovation, Partnerships, and Collaboration: Carbon Storage and Oil and Natural Gas Technologies Review Meeting*, Pittsburgh, Pennsylvania, 2018.

[28] Bakshi, K., "The Seven Steps of Machine Learning," TechLeer, 24 October 2017. [Online]. Available: https://www.techleer.com/articles/379-the-seven-steps-of-machine-learning/. [Accessed 12 September 2018].

[29] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[30] Harlalka, R., "Choosing the Right Machine Learning Algorithm," Hackernoon, June 2018. [Online]. Available: https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f. [Accessed 2018 September 2018].

[31] Google.com, "Machine Learning Crash Course - Training and Test Sets: Splitting Data," Google.com, 22 August 2018. [Online]. Available: https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data. [Accessed 12 September 2018].

[32] Krzyk, K., "Coding Deep Learning for Beginners - Types of Machine Learning," Towards Data Science, Undated. [Online]. Available: https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d. [Accessed 11 September 2018].

[33] Cognub, "Cognitive Computing and Machine Learning," 2016. [Online]. Available: http://www.cognub.com/index.php/cognitive-platform/. [Accessed 11 September 2018].

[34] Mukhopadhyay, S., "Sim-SEQ: A Model Comparison Initiative for Geologic Carbon Sequestration," in *Carbon Storage R&D Project Review Meeting Developing the Technologies and Building the Infrastructure for CO2 Storage*, Pittsburgh, Pennsylvania, 2012.

[35] Kozłowska, M., Brudzinski, M., Friberg, P., Skoumal, R., Baxter, N., and Currie, B. , "Maturity of nearby faults influences seismic hazard from hydraulic fracturing," *Proceedings from the National Academy of Sciences of the United States of America,* vol. 115, no. 8, pp. 1720-1729, 2018.

[36] Obara, K., "Nonvolcanic Deep Tremor Associated with Subduction in Southwest Japan," *Science,* vol. 296, no. 5573, pp. 1679-1681, 2002.

[37] Johnson, P. A., Hulbert, C. L., Rouet-Leduc, B., "Subsurface Stress Criticality Associated with Fluid Injection and Determined Using Machine Learning, U.S. Provisional patent application (not yet filed)". Patent LANS Ref. No. S133652, KS.

[38] Simonini, T., "An introduction to Reinforcement Learning," freeCodeCamp, 30 March 2018. [Online]. Available: https://medium.freecodecamp.org/an-introduction-to-reinforcement-learning-4339519de419. [Accessed 13 September 2018].

[39] Li, H., "Which Machine Learning Algorithm Should I Use?," KDnuggets, June 2017. [Online]. Available: https://www.kdnuggets.com/2017/06/which-machine-learning-algorithm.html. [Accessed 14 September 2018].

[40] Navidi, W., Statistics for Engineers and Scientists, McGraw Hill: New York, New York, 2008.

[41] Kuhn, M., and Johnson, K., Applied Predictive Modeling, Springer, 2013.

Appendix 1. Breakout Session: Resource Recovery and Utilization

Background

The focus questions posed to this breakout group are listed below. These can be treated as the end-points of a study involving DA and ML techniques:

- How do we increase (e.g., double) the recovery efficiency of oil and gas production from unconventional formations?
- How can we improve the ability to track plume migration in the reservoir (following hydraulic fracturing in unconventional wells, and CO_2 injection for EOR associated storage in depleted oil fields)?
- How can completion and stimulation techniques (such as hydraulic fracturing) be optimized?
- How do we optimize the storage and recovery efficiencies for associated carbon storage?
- How can enhanced recovery techniques improve resource recovery efficiency?
- What key signals need to be measured?

A broader set of questions regarding the process of applying DA and ML was identified as a prelude to offering specific R&D topics:

- Do we have the required computing power to tackle these problems? Does the necessary code need to be developed, and personnel trained, to write/support these types of ML problems?
- How are the features that go into the ML process selected? For example, how does one determine the appropriate completion parameters to develop solutions for completion optimization problems?
- Optimizations will be site-specific. Is there a methodology that can be proposed as an overarching strategy?
- How can ML be used to optimize completion and monitoring without interfering with existing optimization?
- How must researchers address the scale, uncertainty, resolution, and environment/economics (SURE) challenge?
- How can researchers improve and employ physics-based models to employ useful synthetic data?
- Could ML be used to tease out more information from time-lapse data?
- How can researchers examine model form uncertainty to build better physics-based models?

Use Cases and Stretch Goals

Several areas were identified where there is a clear need for R&D involving DA and ML toward resource recovery and utilization. The topical areas, as well as the associated key research questions (and the suggested time horizon), are listed below.

1. **Completions and Well Operations Optimization:** (High-Priority R&D Need)
   - Given geology/economics, can DA and ML be used to inform optimum spacing and completion design to maximize value? (Short-Term)
   - How can DA and ML be used with real-time data to improve recovery efficiency (in unconventional reservoirs or storage efficiency in carbon capture, utilization, and storage [CCUS] projects) by automating operations and tailoring completions to site-specific and dynamic conditions? (Short-Term, bigger improvements in the Long-Term)
   - Can DA and ML be used with unstructured data (e.g., field drilling/production reports) to learn from the past and avoid complications and unforeseen situations? (Short-Term, bigger improvements in the Long-Term)
   - How can DA and ML be used for developing next-generation completion design (via reduction/replacement of water use, new proppants, improved chemistry, etc.)? (Long-Term)
   - Can DA and ML help track physical/chemical properties/features that work well in current conditions and have the potential to be optimized with novel approaches? (Long-Term)

2. **Reservoir Management/Operations:** (High-Priority R&D Need)
   - In reservoir characterization, how can DA and ML be used to automate and improve interpretation efficiency for routine tasks such as core analysis, well logging, 3D/4D seismic data analysis, and upscaling of static models? (mostly Short-Term, upscaling challenge could be a Long-Term)
   - How can ML be used with time-lapse seismic, pressure, and/or temperature data to identify/predict saturation front and subsurface fluid movement? (Short-Term for individual data types; more significant improvements with multiple data streams for Long-Term)
   - Can DA and ML be used to identify the optimal EOR strategy following primary production in unconventional wells in addition to CCUS projects? (Short-Term)
   - How can ML be used to tackle the SURE challenge? (Long-Term)
   - How can ML be applied to optimize injection/production in mature fields? This could be for conventional reservoirs as well as for CCUS applications where pressure management is employed using brine production wells. (Short-Term)
   - How can ML be applied to optimize injection/production in emerging (unconventional) fields—using information from previous plays and via transfer learning? (Long-Term)
   - How can scenario-based models for forecasting (incorporating environmental/economic/operational potential changes) be developed? (Long-Term)

3. **Drilling/Geo-steering (for better landing of wells):** (High Priority R&D Need)
    - How can researchers create an automated geo-steering system using DA and ML based on experience in previous wells? The goal would be to reduce time between receipt of information and decision by reducing positional uncertainty (vertical and horizontal) and improving drilling efficiency (rate of penetration). (Short-Term)
    - How can DA and ML be used to identify best locations to drill and the optimal well orientation/spacing/design? (some Short-Term, but mostly Long-Term)

4. **Exploration:** (Low Priority R&D Need)
    - Can DA and ML be used with data from legacy oil and gas operations to determine subsurface heat flux and identify sweet spots for locating geothermal wells? (Short-Term)
    - Currently, DA and ML techniques are routinely used to fill in gaps in well logs, and to create synthetic data. What is the potential for more innovative approaches (e.g., combining well-log signatures, using standard well logs as a proxy for advanced logs, etc.)? (Long-Term)

Data Needs

1. Types

Today's existing data needs consist of (1) drilling data, (2) completion data, and (3) reservoir data, as described below.

a. **Drilling data** includes electronic drilling records (rate of drill bit penetration, revolutions per minute of the drill bit, weight on the drill bit, passive GR emissions, etc.), survey data, wellhead location, measurement-while-drilling/logging-while-drilling (other measurements taken while drilling), cuttings, drilling mud properties, and bottom hole assembly.

b. **Completion data** includes hydraulic fracture stage-level post-job reports (i.e., fluid and proppant volumes, stage/cluster details, treatment pressure history) and re-frac data.

c. **Reservoir data** includes all fluid production (and injection as appropriate) rates/volumes, pressure tests and surveys, open hole logs, time-dependent surface/borehole/cross-well seismic/micro-seismic surveys, distributed acoustic sensing/distributed temperature sensing, diagnostic fracture injection test, surface resistivity surveys, borehole gravity surveys, etc. The ability to visualize different types of data, spanning various scales, on a common platform can present challenges.

2. Availability

While data availability is always an issue, there is a need to catalogue the data currently available in the public domain for potential use in DA and ML-related research projects. In general, data related to drilling and production from existing fields is more readily available.

However, data related to predicting/characterizing undeveloped fields is not as readily available for proprietary reasons. There is also a need to anticipate the data available in the future.

Several sources exist that house public data, including state agencies and websites such as FracFocus, RS Energy, Digital Rocks, Drilling Info, and IHS Markit. A comprehensive data catalogue that can help ascertain what datasets are easily available from such sources, and thus increase broader access to important datasets, would be a valuable resource for the research community.

A significant portion of intellectual property is developed within oil and gas companies and is unpublished. This limits the amount of data available for developing DA and ML-based models. DOE does not have large amounts of data available from any single play. However, there are DOE projects that may be valuable sources of data as a starting point for any given play. It could be beneficial to solicit industry to contribute data to add to the DOE data on a play-specific basis (via a Joint Industry database type arrangement) that could become a rich source for future research projects.

3. Opportunities

It is important to ascertain what data types and datasets operators are willing to share with DOE and/or place in the public domain. Derivative datasets from operators are possible. However, a clear case needs to be made regarding the potential use of data for DA and ML-related research projects, and the benefits that would be derived by companies contributing data. In this context, it is useful to note that data consortia for both conventional and unconventional reservoirs have been put in place by companies such as Core Labs, which could be a useful data-sharing model going forward.

4. Strategies

One compelling argument for companies to share their data is that doing so allows for the greater academic/governmental community to approach problems faced by the company doing the data sharing (i.e., “open-sourcing” the problem). This situation is only feasible for companies having little existing competition in a particular play.

There is a heavy economic pressure for operators to share their data with a broader community, with the objective of attracting greater intellectual capital to open-source complex problems. This is particularly effective in plays where there is little competition for new leases (operators have moved into manufacturing-style development). This allows cost-effective solutions to be developed for problems that benefit the entire industry, without harming any one company’s competitive advantage.

5. Data Formats and Integration Issues

There is a need for standardized data formatting. To facilitate standardization, there is also a need for long-term partnerships between DOE and producers. This will enable DOE to identify valuable data for enhancing technology that producers have not identified. Additionally, there are efforts at a non-profit organization called Digital Rocks to standardize the format of available data to enable use for ML.

A discussion was had concerning DOE's role in data collection/sharing for the purposes of using it for ML applications. Producers posed that they have data among themselves and have hired (or are hiring) data scientists to leverage it. If all the necessary data exists today, DOE has no role to play in the data aggregation space. However, this is not entirely the case. If there is useful data that the producers are unaware of that can advance the technology, it can be DOE's role to fill the gap. It was agreed that DOE's primary role is to enable identification of these types of value-adding data types to advance the current state of the technology. To support this function, long-term partnerships between DOE and producers are far more useful than short-term.

Other Topic-Specific Information

1. Barriers/Missing Data

- **Completion and Well Operations Optimization:** The cost associated with acquiring data means that operators typically collect the bare minimum of data to sustain operations. The overall quality of data is low, since there is typically surface data and not data at depth, where stages are isolated. For these reasons, there is often not enough data to use in ML analyses to improve completion performance. However, given an appropriate data storage substrate, it is possible to generate synthetic data. For example, operators could train neural networks on real data at a well-monitored site or utilize high-fidelity simulators to generate synthetic data and then turn those data over to collaborators. Some suggested research topics regarding this subject are

  i. Use of a suitable model to generate synthetic data
  ii. Identification of additional data needed by determining the value that data can add
  iii. Identification of cost-effective sensor(s) that can be applied at all wells

- **Reservoir Management/Operations:** As previously mentioned, there are not enough complete datasets from public sources to acquire the specific and detailed level of information to diligently investigate all of the R&D challenges presented throughout this document. Putting non-disclosure agreements (NDAs) in place may facilitate public and private companies to share issues, data, and key findings while protecting intellectual property rights. Exclusivity to work on a study with DOE or the national laboratories within a given basin should not be granted to any oil company operator.

  There is a need for one or more universal datasets that can serve as benchmarks as DA and ML become mainstream technologies. For many years, the Society of Petroleum Engineers had a comparative problem-solving challenge in reservoir simulation using standardized datasets for a variety of problems. The National Energy Technology Laboratory's Carbon Storage program performed a similar exercise called Sim-SEQ, where multiple teams from different organizations conducted a model comparison initiative involving both model-to-data and model-to-model comparison at a common CO_2 storage field site [34]. A similar approach could be adopted, with DOE taking the lead in formulating several DA and ML application problems, selecting the appropriate dataset with industry collaboration, and collating the solutions as a knowledge sharing exercise.

Another challenge here is that insufficient quantities (and perhaps quality) of data being collected in-between wells (pressure, state of stress, porosity, permeability, chemistry) can be an issue. Can DA and ML-assisted model validation studies be helpful in this regard? Are there remote sensing technologies that can be used in conjunction with DA and ML approaches for such problems?

Currently, proprietary data formats are a barrier to entry, which potentially could be overcome through the use of a third party. It is also important to develop pathways to manage the data that exists, including compression, visualization, and standardization.

Some regulatory agencies are also barriers to data collection and sharing because they prescribe data types that should be collected and shared in the public domain. A common data standard, adopted across multiple state jurisdictions, would make it easier for companies to collaborate in plays crossing several state lines.

- **Drilling/Geo-steering:** More data should be collected to ensure that with the well surveys, the positional uncertainty along the wellbore is better quantified. This may enable improved targeting of so-called "sweet spots," with higher porosity, total organic carbon, higher density of natural fractures, and more brittle rock.

- **Missing Data:** There is a need to measure/estimate properties between wells (e.g., pressure, stress, permeability, minerology, chemistry). This could be done by improving the resolution of seismic surveys, integrating multiple data streams (to generate surrogates for missing data), etc. Potentially, DA and ML could play a role by combining physics-based models with data-rich signals that can help infer characteristics of unsampled regions.

  The use of DA and ML to facilitate integration of models that represent different physical processes (e.g., mechanics/fracking, production, kinetics) also has interesting potential with respect to both model validation and development of proxy (surrogate) models.

  A big challenge in data-driven modeling is the role of overlooked or missed data. Thus, there is a need to identify such data by developing methods for automated data discovery or smart searches in the public domain. This could include both structured (numeric) and unstructured (text, image) data types.

Often, the frequency of data collection is irregular or sparse (both in space and time) and hinders the development of improved-resolution models. To this end, DA, ML, and multivariate correlation techniques could be applied to generate denser sets of data.

A meta-analysis of different types of data could be carried out to determine why more of certain data types don't exist (e.g., cost, lack of technology). This would involve a combination of data mining and abstraction of subject matter expertise.

There has been a greater proliferation of downhole sensors, which provides the opportunity to build data-driven models that could offer better ways to accurately convert surface data to downhole data. A related challenge is the possibility of starting with only a limited set of dense downhole measurements and converting them into additional parameters.

2. Tools and Techniques

- **Data organization/management:** The use of open-source and platform agnostic approaches would likely contribute to better data organization and management. In line with FAIR principles, it is unclear if existing platforms are sufficient or if industry needs application-specific capabilities.

  With the help of professional societies, industry could establish standardized data formats. Industry would also benefit from the application of graph databases to accelerate the data wrangling process.

  There is a need to better describe the ontology of the data being created (using a framework such as the application program interface standard for ontology).

- **Analytics/knowledge discovery:** Physics-based models could be used to effectively constrain open-source ML models. Another interesting possibility is the development of novel constitutive models using general DA and ML approaches, perhaps to enable production from more clay-rich ductile reservoirs industry has struggled to produce or has simply dismissed as not being easy to hydraulically fracture and prop open.

  The use of ensemble methods can add to modeling confidence. This involves incorporating multiple DA and ML approaches and aggregating the results, as opposed to using a single technique. Some open-source applications/tools are already available for this purpose.

  Fusing disparate data streams could provide additional insights. Some financial service companies have extensive experience with non-stationary time series data and could provide best practice examples.

  One approach to involve Silicon Valley startups in mainstreaming oil and gas DA and ML applications would be for the industry to provide synthetic data generating models. This would remove the burden of having NDA agreements that generally are an impediment. The availability of anonymized datasets or artificial streaming of sensors for others to test tools would help in testing the applicability of new and/or emerging technologies that have not been applied in the oil industry.

Efforts are needed to promote moving from correlation-based models to causation-based models. The use of Bayesian Networks has been a starting point, but more needs to be done on this topic.

Knowledge abstraction from “soft” (i.e., unstructured) data remains a relatively unexplored topic in oil and gas DA and ML applications. A key imperative will be to combine written logs and subject matter expertise to abstract the domain-specific knowledge from experienced professionals into a useful form that can help train the next generation of engineers/scientists and to automate routine tasks. Can crowdsourcing play a role?

- **Decision support/visualization:** It is important to validate models and convey information with visualizations. To enhance current visualizations, oil and gas companies should involve tools/learnings from the entertainment and gaming industry. There are also many open-source game engines available for use in this space.

Appendix 2. Breakout Session: Autonomous Monitoring

“Smart wells” and “intelligent fields” have been a stated goal of many subsurface operations (notably oil and gas operations) for several years, but the promise of such systems has not yet been fully realized in the field. With recent improvements in the resolution and accuracy of monitoring systems, as well as recent advancements in DA and reservoir simulation, the real-time use of downhole and other sensor data to inform operational decisions is close to becoming a reality. A key enabler of intelligent fields is the development of autonomous monitoring platforms, which consist of sensors coupled with ML-based algorithms that convert data to knowledge. In resource recovery applications, autonomous monitoring can enable new options for reservoir operations with improved efficiency, recovery, and safety. In carbon storage applications, autonomous monitoring can help in tracking the carbon dioxide and pressure plumes over time, identify signs of a wellbore or seal leak, and locate signs of potential induced seismicity. Overall, the time is opportune for the realization of autonomous monitoring systems that could transform approaches to subsurface energy systems.

This section presents an assessment of opportunities and needs related to autonomous monitoring in the subsurface, as identified by an expert group spanning DA subject matter expertise in the subsurface, as drawn from the hydrocarbon industry (majors and independents), the service industry, national laboratories, and academia. In addition, this group reflected expertise in conventional reservoirs, unconventional reservoirs, CO_2 storage, and geothermal systems.

Background

Autonomous platforms—i.e., based on the coupling sensors to ML algorithms that extract and analyze signals—are transforming the ability to gain critical knowledge about subsurface environments. Autonomous platforms can reduce monitoring costs significantly and speed up the path from data to knowledge by removing protracted and laborious analysis. In addition, machine-based interpretation of data can reduce bias in signal interpretation, leading both to the identification of new signatures and to the extraction of knowledge from noisy signals. ML algorithms can handle continuous data from sensors, opening a wider use of data, and can incorporate multiple datasets in an analysis, leading to a deeper understanding.

Consequently, autonomous monitoring is an enabling technology for many future subsurface operations. In long-term CO_2 storage, post-injection site care costs can double the cost of an operation based on conventional monitoring platforms, but autonomous monitoring could provide cost-effective solutions to ensure that a site is performing well. In unconventional reservoirs, only a small fraction of the original hydrocarbon in place is recovered using current reservoir management strategies, but autonomous monitoring could provide the rapid insights into reservoir behavior that are needed to optimize recovery. Finally, in all subsurface operations, autonomous monitoring can lead to increased safety and decreased probability of unwanted events by providing real-time information on critical changes in a subsurface system ranging from the state of stress to fluid pressures.

Autonomous monitoring involves the conversion of data to knowledge via analysis and visualization. Near-term goals for autonomous monitoring focus largely on improving decisions made by a person (i.e., human-in-the-loop), but long-term goals include fully automated systems (i.e., autonomous monitoring and control).

Autonomous platforms range from those that function at the edge (of the network) to those that operate in the cloud. In an edge configuration, data from the sensor are analyzed at the collection point, typically resulting in either a decision or in a winnowing of the full spectrum of the continuous data into a smaller set of key information. In a cloud configuration, analysis is done on data that have been transferred from the sensor array to centralized storage where it can be analyzed in full, perhaps in conjunction with other datasets.

Physics-based ML adds a critical dimension to autonomous monitoring for subsurface systems (Figure 9). Specifically, the ability to predict the behavior of and signals associated with a wide range of processes helps to address the generally limited availability of labeled datasets from real, natural systems as needed to apply ML algorithms. Data availability, which is discussed in greater detail below, is only one root cause of the limitation in labeled datasets for subsurface systems. Two other root causes plague all subsurface operations. First, the complexity and variability of subsurface systems is rarely probed at sufficient detail to provide many of the labels that might be needed, leaving imprecise empirical relationships with significant uncertainty; the system characteristics used in generating physics-based synthetic data, however, can be completely captured in the labels, allowing machine-based algorithms to learn more precise empirical relationships and, potentially, enabling the extraction of information from the empirical scatter. Second, subsurface operations are expensive and must meet permitting requirements, inhibiting the ability to establish field-test sites for many scenarios of interest; thus, limited access to real field data exists for these scenarios, which is needed for discovering signatures to monitor and for developing and testing autonomous systems.

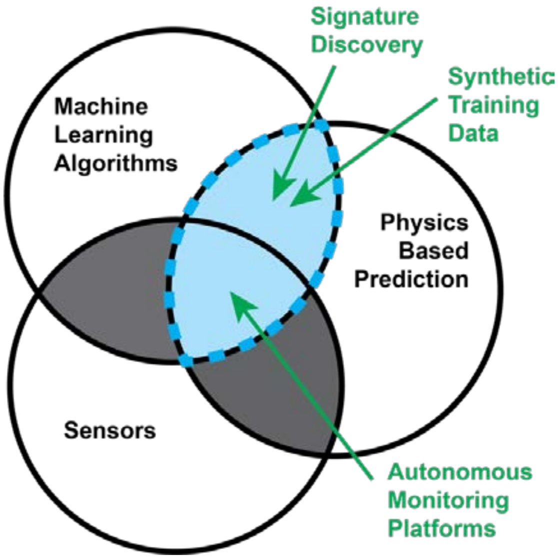

Figure 9. Venn diagram showing opportunities for physics-based ML in the context of autonomous monitoring for subsurface operations. Coupling of physics-based prediction and ML algorithms can lead to a virtual learning environment for discovering new signatures to monitor system behavior and for (pre-)training ML algorithms prior to field deployment. Coupling of physics-based ML with sensors can lead to new autonomous monitoring platforms that speed the conversion of data to knowledge as needed for various decisions.

Use Cases

The identification and consideration of use cases helps in evaluating the needs and opportunities for autonomous monitoring. Several categories emerge, based on the time scale of the decisions that need to be made.

Use-Case Category	Time-Scale of Decision	Drivers for Decision
Safety	Seconds to Hours	Protection of Workers, Public, Equipment, Environment
Wellhead Operations	Hours to Days	Drilling Optimization • Minimizing bit wear • Maximizing rate of penetration • Controlling bit trajectory/position • Logging while drilling Well-Operations Optimization • Optimizing production rate for recovery • Optimizing injection rate for controlling sweep efficiency, pressure management, etc.
Reservoir Management	Days to Months	Resource Recovery/Utilization • Designing strategies for secondary/tertiary recovery, re-stimulation, infill drilling, etc.

Use-Case Category	Time-Scale of Decision	Drivers for Decision
		• Controlling/exploiting evolution of reservoir state of stress Long-term Monitoring • Ensuring containment (no leakage) • Demonstrating conformance in reservoir performance Avoiding Induced Seismicity • Optimizing rate of extracting co-produced fluids • Identifying changes in state of stress outside of reservoir

Table 2. Autonomous monitoring targets fall into three use-case categories based on the time scales of decisions.

Safety decisions often have an immediate need for a decision, perhaps on the order of seconds to hours. Examples of safety-related issues include well blowouts, containment failure (at the wellbore), and cyber-attacks. Autonomous monitoring platforms offer the potential to address a range of safety-related research needs:

- Better/new monitoring systems to detect conditions of concern during drilling
- Data integration to detect precursor events, sequences of concern, and patterns of failure
- Autonomous monitoring based on an internet of things

Research tied to this use case must also consider how an autonomous monitoring system interfaces with decisions, particularly for those scenarios that involve human-in-the-loop. Finally, safety-related applications will require demonstrating that an autonomous monitoring platform is sufficiently reliable, which could include a research strategy for certifying such platforms.

Wellhead operations decisions often have a need for a quick decision, perhaps on the order of hours to days. Examples of decisions-at-the-well include real-time drilling-and-completion operations (optimization of drilling dynamics, steering, and optimization of stimulation) and reservoir-management operations (optimization of production and recovery). Autonomous monitoring platforms offer the potential to address a range of decisions-at-the-well research needs:

- Rapid data integration to elucidate subsurface conditions and behavior
- Better/new monitoring systems, especially for conditions far from well
- Higher value of information
    - Data to knowledge that maximizes user understanding and confidence in results

    - Dynamic identification of the types and timing of new data needed to improve understanding by lowering uncertainty

Reservoir-management decisions can often be made on a longer time scale, perhaps on the order of days to months. Examples of reservoir-management decisions can occur both during active reservoir operations (e.g., modifying injection/extraction rates to optimize recovery/sweep and to minimize potential for induced seismicity) and following active reservoir operations (e.g., during a post-injection site care phase for long-term storage of CO_2). Autonomous monitoring platforms offer the potential to address a range of reservoir-management research needs:

- Rapid data integration to forecast subsurface conditions and behavior
- Dynamic coupling of autonomous monitoring with autonomous updating of predictive models for improved forecasting
- Signal processing—to speed analysis and to remove bias
- Uncertainty quantification

Stretch Goals for Autonomous Monitoring R&D

In consideration of the needs/potential summarized above, stretch goals were identified for autonomous monitoring R&D for each use case. Both short-term (3 years) and long-term (10 years) goals were considered.

Safety: Development and demonstration of platforms for real-time detection of anomalies (or sequences of concern) for health, safety, and environment purposes with high (near 100%) reliability

> *An area of particular need/opportunity for autonomous systems is safety associated with failures at a well, including*

- Prevention of blow-outs and loss of well-pressure control during drilling
- Prevention of wellbore failure in pressurized reservoirs (e.g., gas storage reservoirs)

Wellhead Operations: Development of platforms for (1) enabling “manage by exception” for subsurface reservoirs and (2) real-time conversion of data to knowledge for field operations

> *Areas of need/opportunity for autonomous systems associated with decisions at a well include*

- Real-time updating of well data (e.g., pressure, volumes/compositions of produced fluids, etc.) with near-term forecasting of reservoir behavior (as needed to manage injection/production rates for optimized recovery)
- Real-time interpretation of geophysical data (e.g., microseismic data) as needed to manage the stimulation operations to optimize subsequent production from an unconventional reservoir

Reservoir Management: Discovery of new signatures (and associated autonomous monitoring systems) for (1) monitoring the subsurface far-afield from the well and (2) monitoring the subsurface outside of the reservoir

Specific areas of need include

- Identification of (changes in) the fluid/pressure/stress distribution within a reservoir to improve sweep efficiency, recovery efficiency, etc.
- Identification of changes in state of stress outside of a reservoir to avoid induced seismic events
- Identification of leakage outside of the primary reservoir to ensure containment of a stored fluid/gas
- Identification of long-term wellbore degradation and aging (both casing and cement) that foretell a weakening of primary or secondary well containment

Data Availability

Data availability is central to ML, with many challenges that are common across the subsurface research space. Several challenges, however, are worthy of specific consideration in the context of autonomous monitoring:

- **Failure-related data** are particularly important for training of safety-related autonomous systems, and these data may have limited availability. These data could include, for example, signals associated with an event of concern labeled appropriately relative to the details (e.g., timing, location) of the event of concern. In addition, these data include information on events/conditions that are precursors to failure. There is a particular challenge associated with training data needed for low-probability but high-consequence events, which are unlikely to be observed in high enough frequency to train a ML algorithm with high enough accuracy for reliable safety decisions. One strategy may be to develop algorithms that incorporate other types of prior knowledge (e.g., regional geology) and that have the capability to learn/adapt from data over time.
- **Data are needed for signature discovery and training/testing of autonomous monitoring platforms**, yet data are limited for some applications.
    - *Data on unconventional reservoirs:* Data on the behavior of unconventional reservoirs are limited in comparison with conventional reservoirs, in part due to the wide range in site-specific properties that can impact behavior and, in part, due to the proprietary nature of many datasets. An operator may have sufficient data to train an autonomous monitoring system for conventional reservoirs, using a combination of owned proprietary data and the wealth of publicly available data amassed over decades. However, operators are unlikely to have sufficient data to train an autonomous monitoring system for unconventional reservoirs based solely on their own proprietary datasets and on the limited publicly available information.

- *Data on subsurface systems outside the reservoir:* For example, data on leakage impacts outside a reservoir are needed to train an autonomous system for post-injection monitoring for CO_2. By design, storage operations are intended to be devoid of leaks, so they do not result in datasets that can be used for training.
- *Data on well integrity (static and dynamic)*: This type of information is not consistently collected and/or available.

A strategy for addressing the need for failure-related data could include public partnerships with industry, perhaps using a consortium model (e.g., joint industry project) to pool data, thereby achieving larger datasets and minimizing any potential risk to individual entities associated with sharing sensitive data. However, the need for a relevant field site to demonstrate and certify autonomous systems used for safety applications will likely necessitate the development of a test site/facility that can produce engineered failures.

The use of synthetic data could be a powerful option for autonomous monitoring needs, both for signature discovery and training datasets, and particularly for unconventional reservoirs and subsurface systems outside of the primary reservoir. The coupling of physics-based simulators with empirical models (e.g., ML algorithms and other ROMs) could be exploited to capture the behavior of even complex subsurface systems, generating a virtual subsurface environment that could subsequently be coupled with physics-based platforms to simulate relevant potential signals. This approach relies on physics-based simulators that have been adequately validated and calibrated with real data to ensure a high quality of the resulting synthetic data. Capturing full physics of the system is essential for this approach, which can pose a computational challenge for generating sufficient realizations to build the virtual reality environment; however, ML algorithms can be developed to identify strategic sets of high-fidelity simulations needed to improve the synthetic dataset for training and to elucidate artifacts in synthetic data generated by ROMs relative to deterministically derived synthetic data. Once a sufficient dataset is generated for potential signatures, ML algorithms can be used to explore the virtual environment for signatures or utilize it for training of an algorithm that will be incorporated into an autonomous platform. The incorporation of synthetic noise can at least partially mimic the challenge of detecting signals in variable and noisy natural environments.

The use of data from real sites is always preferable to synthetic datasets for both signature discovery and (final) training of autonomous systems, inasmuch as they embody the complete physics of the system as well as site-specific characteristics (including the natural variability and complexities). However, it can be useful to initiate efforts with synthetic data, which are less susceptible to challenges associated with uncertainties in ground truth and which can be generated to produce a range of signal/noise ratios. Efforts can then transition from synthetic to real datasets.

The coupling of real and synthetic datasets (i.e., data fusion) can be incorporated as a strategy in data analysis, addressing limitations in each type. Data fusion will require addressing several technical challenges, including temporal synchronization, variability in data quality/relevance, etc.

An additional strategy for signature discovery and training/testing data is the use of analog sites. As an example, some natural CO_2 reservoirs may offer conditions to study signatures associated with leakages.

Legacy data also offer a potential opportunity for training and testing of autonomous systems. In many cases, however, these datasets may require additional investments in DM to archive (and consolidate) datasets and to convert data into standardized digital formats that can be used by the research community. This poses a challenge of handling disparate types of file formats ranging from non-digitized data formats (e.g., images, well logs, seismic data) to information embedded or embodied in text form (e.g., in reports). These challenges present an additional ML opportunity.

Finally, some autonomous systems may require development of engineered field sites that can provide the conditions necessary to develop and test platforms.

Tools and Techniques

Several existing tools and techniques can be adapted for use in autonomous monitoring applications:

1. **Coupled physics-based and empirical-based learning** to train and evolve autonomous platforms
    - Physics-inspired ML
        - Using physics-based models to train ML algorithms
        - In autonomous monitoring, this includes end-to-end multiphysics models (source to signal)
    - ML embedded models
        - Can be used to speed physics-base simulations

2. **Visualization platforms (including augmented reality platforms),** such as for
    - Data (attributes); Topology
    - Model Parameters
    - Predictions/Outcome/Consequences (”regrets”)
    - Confidence/Uncertainty (including value of information)
    - Continuous datasets; large volumes of data; multi-dimensional (>3D)
    - Multi-variate, spatial detection/visualization of outliers/anomalies

Effective visualization of information from autonomous monitoring systems will be critical for human-in-the-loop applications. Although tools exist, they may require investments in adaptations as needed for subsurface applications, specific types of decision-makers, etc.

3. **Data compression,** particularly for continuous data streams. Specific additional needs for autonomous monitoring include
   - Handling, visualizing, and curating large amounts of data

4. **Forward deployment** should be a key stated goal for R&D efforts in autonomous monitoring, emphasizing the need for such systems to be economically deployable in the relevant field scenarios and conditions.
   - *Raspberry pi* is one recent technology that could be exploited by autonomous systems for the subsurface, providing a low-power, portable (credit-card sized) platform to couple ML algorithms to analyze sensor data at the edge.

Barriers and Challenges

Several potential barriers and challenges should be considered in the development of autonomous systems for subsurface operations:

- **Adoption of new technology by decision-makers in their decision processes:** Cultural biases may exist among different groups based on differences in personal experiences on how to approach a problem (e.g., deterministically vs. empirically vs. probabilistically), based on comfort-level with new technology, based on educational traditions, etc. R&D efforts to design autonomous systems consider these factors from outset to ensure
  1) effectiveness of integration of new platforms into decision-workflow
  2) understanding of need/value of data collection and data curation

  Autonomous systems must fit within the decision workflows established for specific applications.

- **Limitations on data volumes (e.g., bandwidth, or data storage, as needed to handle continuous data streams from sensor arrays) and on required response time for information:** Utilization of edge computing (e.g., via Raspberry pi platforms) is one possible strategy to address these barriers, by pre-processing continuous data at the source to produce a smaller volume of derivative data of interest. This pre-processing can be designed to produce key information as needed by decision-makers, thereby speeding the timeline of data collection to knowledge. But, the effective use of edge computing will need to overcome several technical challenges:
  - Ability to migrate new algorithms from the cloud (where they are developed) to the edge (where they are needed for continuous improvement). This capability would allow edge-based algorithms to remain current as new information is developed, for example, on the nature of a specific site or as new signals are identified.
  - Ability to integrate combine distributed analysis at edge in conjunction with an individual sensor with integrated analysis at cloud, which can incorporate information from multiple sensors and other dynamic site-specific information.

    - Ability to accommodate the heterogeneity of various hardware/software platforms that exist at a site.

- **Lack of appropriate or useful signals:** This relates to (1) signals that have not yet been identified, and (2) small signal-to-noise ratios:
    - For example, data collection is often focused on signals relevant to success, but some use cases (e.g., safety) require data/signals on failures and data/signals on rare events. As noted above under Data Availability, strategy is needed to address this gap.
    - For small signals buried in noise, ML algorithms could be developed to improve extracting signals from existing data/sensors. In addition, the use of physics-based models can be used to discover new signatures that may be buried in existing datasets, thereby guiding the development of new autonomous platforms.

- **Need for database management protocols:**
    - Databases must be maintained current at a sufficient refresh rate to allow autonomous monitoring/analysis in the cloud.
    - A chain of custody must be established, including the identification of any manipulation (preprocessing) of data that has occurred prior to cloud/storage.
    - Metadata and other non-numerical data need to be archived and made available in a useable form as needed in training autonomous systems and/or as might be needed in autonomous analysis. These data types could include non-numerical data from reports, operations, logs, etc.

Appendix 3. Breakout Session: Seismicity and Dynamic Stress States

A key question for which more definitive answers are needed is how to determine the physical and mechanical mechanisms that allow induced seismicity to occur. It has been generally accepted that changes (typically increases) in pore pressure due to injection (or extraction) are responsible for changes in effective stress that lead to fault slip. However, there are likely other physical mechanisms such as poroelastic stress changes that can influence fault slip as well. In addition, the pervasiveness and configuration of faults and rock lithology that allow for seismic slip likely control the distribution of induced seismicity. Fortunately, there is a growing number of publicly available datasets and improving geomechanical simulation strategies that can help to test the proposed mechanisms and discern the primary factors. Continuing to improve the collective knowledge of rock rheologies and the imaging of faults in the subsurface should ultimately lead to a better understanding of which faults represent higher hazard.

Ultimately, the goal is to enable forecasting of induced seismicity, such that the necessary data, signals, and processing workflows are considered. Seismic recordings are critical, so a transparent public seismic network is necessary, but unfortunately, data are currently much sparser in the target regions of the Central and Eastern U.S. than in areas like California. Improving seismic networks will not only improve event detection, but new imaging techniques can utilize them to look at changes in reservoirs over time. Publicly available active seismic datasets are scarcer, but will be critical for analyzing microseismicity patterns, and can also be used to construct high-resolution images of the reservoir. Beyond seismic data, the most pressing data need is for pressure monitoring, both in the near field and far field, and preferably downhole. Additional monitoring and estimation of poroelastic stresses (i.e., deformation relative to pressure applied) is also needed. There is also a need to incorporate more information from both natural system, land use, and operation/engineering. Other datasets of interest include structural geology (for faults), geochemistry, electromagnetic, and lab measurements. This creates a rather long list of data and signal types that would be beneficial to have on hand, but ML approaches can help to understand the relative importance of the different contributions and prioritize future efforts.

In all cases of data that vary over time, the more continuous the time series collected, the better. The desire for more critical data highlights the needs for improved data sharing by targeting key new data collection efforts and promoting open data access. There will also be increasing DM needs that require better management systems and promoting uniform data standards. A variety of new processing workflows will likely be developed to analyze these increasing datasets, so there will be a corresponding need to apply statistical approaches to the observations, simulations, and models to assess confidence in the new information.

With improved data and processing workflows in hand, one can consider strategies for how to implement forecasting of induced seismicity during injection and production operations. While seismicity directly induced by pore pressure changes from injection/extraction is the focus of this development, forecasting will ultimately need to consider triggering relationships between earthquakes as well. Utilization of seismic data for forecasting shows promise (see Case Study 1 below), but it is important to have a two-pronged approach that incorporates both enhancing traditional seismicity catalogs and processing the continuous seismogram data to look for new clues of fault behavior. ML has been successfully applied to both approaches but needs to be applied to a broader set of data types as the success will depend on which data types are available and incorporated. For example, the impacts of production and extraction operations need to be considered, not just for injection operations. This also highlights the need to develop geomechanical models via simulations that feed back into forecasting. Real-time forecasting is the near-term goal, but the more transformational goal is to forecast before injection or production at a proposed well. Further evaluation of the regions that are not experiencing induced seismicity despite operation activity is key to improving the assessment of the relative hazards.

As the capability to identify and forecast the induced seismicity hazards improves, it will be necessary to consider how to manage these hazards and reduce the risks, seeking the long-term goal of avoiding any damaging seismicity. A component of this is recognizing that different areas have different risks even if hazards are similar (e.g., offshore vs. onshore operations). The central strategy is via pore pressure management, but this can be separated into a variety of approaches: (1) reducing the amount of fluid being disposed, (2) changing the rate of injection/extraction, (3) adjusting where the pore pressures are being changed, (4) changing the viscosity of fluid injection (i.e., fluid engineering), and (5) improving understanding of the properties of the injection interval. These approaches all need further evaluation to improve management strategies. Yet one also needs to consider what influences the risks, including what amount of time matters in the estimation of hazard, how much the duration of operations influences what is induced, and how much time is needed for an efficient response (likely depends on type of operation).

Primary Targets for Investigation

From the primary issues associated with induced seismicity and understanding dynamic changes in the stress state of the subsurface, the research community can work on several near-term and transformational targets that have been identified. The primary near-term target identified was to provide a replacement for the Traffic Light Protocol (TLP), which is a common strategy to manage induced seismicity. This can be replaced by a more nuanced understanding of hazard and ramifications/risk that is properly placed in the context of how the situation is developing over time. In essence, the goal is progress toward operators being able to make changes to their operations in real-time to limit/reduce/prevent induced seismicity. This will likely require improved utilization of seismic recording to establish reliable statistical likelihoods of earthquakes of given magnitudes. It will also require progress on developing a pore pressure management system that involves improved measurement of downhole pressures, understanding of how changing viscosity of injected fluid changes the likelihood of seismicity, and strategies for reducing injected fluid (i.e., improved recycling) volumes. Applying ML to a broader set of data types can improve forecasting. Currently, there is a hope that ML will provide improved understanding of key variables (i.e., which data types are most important for forecasting) based on preliminary work that demonstrates ML can provide new insight into factors leading to induced seismicity. However, there is a need to clarify which aspects of ML are well designed to tackle this problem, and which datasets are optimally designed for incorporation. Ultimately, in transitioning from a TLP to a real-time forecasting approach, the regulators need to be included more frequently in the scientific discussion. Communication will be crucial to keep forecasting technology in line with policy and applicability. There are opportunities to refine regulatory policies to be efficient, but studies need to be well vetted to convince regulators that new technology is trustworthy. Even something as basic as deciding on the regulatory thresholds can be improved by comparing societal risks with scientific discernment. Moreover, social scientists and economists can be brought into the discussion to better understand the impacts on society and how to mitigate risks.

The long-term transformational targets for future investigation can be loosely collected under the idea of engineering the subsurface to manage induced seismicity. The ultimate goal would be to entirely prevent damaging seismicity, but there are a variety of issues to investigate on the pathway to this goal. These would include the potential for autonomous long-term monitoring that could be used to provide feedback into a controlled engineering system for the operator. It would also include identification of key changes to make before even starting operations to avoid/prevent seismicity from reaching a regulatory threshold. Improving physics-based models of forecasting would likely require better understand of earthquake physics, the initial subsurface stress state, and how stress states change dynamically. Regardless of how the forecasting strategy develops, a long-term goal would be to consider practicality in the science and eventual policy to ensure that regulators can implement findings. Moreover, gains in understanding of these processes in the subsurface would also have the potential to improve the development of reservoirs. For example, these developments would present opportunities for improved reservoir characterization that could ultimately improve recovery and production.

Data Availability

Achieving success for both near-term and transformational targets will undoubtedly be influenced by data availability. Of the currently available types of data, foremost is continuous seismic data recorded at the surface, sometimes in small local networks around operations (<10 km spacing), but often at the regional scale (~100 km spacing). Some operational data are currently available, including injection volumes and pressures, as well as production data. Injection data are often reported monthly to regulators but may also be reported daily and occasionally when there are higher sample rate times series. Production data are typically quarterly or annual. Geologic data are often available, primarily in well logs or as core. Remote sensing (e.g., InSAR) and potential field data are often available but rarely utilized. A variety of surface datasets are available, such as those related to hydrology, climate, and land use, which could be useful in investigations of induced seismicity. Finally, a growing amount of laboratory data are available and relevant to this problem, including continuous seismic and deformation recordings, direct measurements of slip, shear stress, friction, and a larger control on the injection parameters.

Another important consideration relates to the types of data that are most desired, though it is important to consider the steps necessary to accessing this data. The breakout group sought to construct a list based on a perceived order of priority, but this is certainly influenced by the makeup of the breakout group and may also reflect longer-term prospects being perceived as lower priority. Foremost, there is a definitive need for more continuous pressure data, with downhole measurements better than those at the surface. There is also a need for better monitoring and estimates of poroelastic stress changes. In terms of seismic data, there is a need for more downhole continuous recordings, with DAS being a promising technology, to improve detection of microseismic sources and imaging of the structures. More detailed reporting of injection and production histories is needed, including full stimulation reports. Improved and expanded fault data are important, even if it is unclear yet how best to evaluate the risk of any given fault identified. A more complete dataset of geologic/lithologic properties downhole would provide opportunities for improved understanding of the rock types involved in induced seismicity. Laboratory datasets can also be improved through more publicly available continuous recordings, estimates of velocity across fault zones and how it changes, and other estimates of dynamic moduli. In situ measurements of hydrologic parameters, such as permeability and diffusivity, chemical variations such as those involved in reactive flow, and direct measurement of slip including displacement and fractures can also be improved. Downhole geodetic data, such as that from tiltmeters and decision support systems, could substantially improve the understanding of aseismic deformation. Simulation datasets are also important and can be particularly helpful in developing a greater understanding of faults, fracture, and poroelastic behavior. For example, one could use ML to optimize the simulations when comparing with lab data, and the statistical features and insights from ML can be scale invariant. While not currently utilized, downhole electromagnetic sensing could be very useful for monitoring dynamic changes associated with these processes. InSAR data has only been used on occasion to investigate induced seismicity and could be used more extensively to assess the impacts of injection and production on induced seismicity.

Opportunities for ML to Help Achieve Goals

As discussed in previous sections, ML can help to reach the key targets for dealing with induced seismicity (primarily preventing damaging seismicity) through enhancements such as faster processing, reducing bias, integrating larger and disparate datasets, and dealing with high-dimensionality. This section highlights some specifics of how ML can help overcome key obstacles.

The breakout group identified replacing the current TLP strategy (e.g., if mag>X, then do Y across a broad region/state) as a primary need for dealing with induced seismicity. The broader research community is seeking a more nuanced understanding of hazard during the operational phase that is properly placed in the context of how the situation is developing over time. The goal is an evolving estimate of the likelihood of a specific scenario (certain size event, certain amount of seismicity) depending on what the operator or regulator chooses. Over the short-term, improvements can be made by employing forecasting using improved microseismicity detection along with continuous waveform approaches. ML can improve seismic event detection in real-time through faster processing and enhanced clustering analysis. Moreover, ML can deal with large continuous waveform datasets to look for changing patterns over time to anticipate bigger events (see Case Study 2 below).

Over the longer term, the TLP replacement should move toward a probabilistic estimate of how companies should inject/produce at a given well that is time dependent by combining probabilities from the likelihoods derived from different datasets. This approach would seek to integrate site-related data and process-related data and a range of temporal input types (static vs. dynamic) to help embed the physical processes in the estimation. The potential types of information that would be integrated are hypocentral distributions, seismicity rates, frequency-magnitude patterns, long-term slip rates, short-term stressing rates, estimates of forcing from injection and production data, and lithologic and hydrologic data. ML would be needed to optimize the integration and evaluate the contributions from the different datasets to incrementally improve the forecasts over time. ML can also be used to improve the information from each data type, from basic statistical approaches up to topological clustering; deep learning should also be considered to help deal with high dimensionality and reduce bias. Ultimately, a trust metric will be needed to evaluate the resulting likelihood model. For example, a trust metric from deep learning could identify if the model output is not as reliable as it should be and how well the model utilizes the input space. Convolutional neural networks, committee machines, or adversarial networks (distributed intelligence) could potentially be used to evaluate outputs. These efforts are not regularly undertaken now but will be important to implement as either the operator or the regulator.

A different goal where ML can also be applied is to better understand the physical mechanisms leading to induced seismicity, particularly as this could lead to improved understanding of natural seismic hazards as well. For example, a better understanding of the relationship between stress and induced seismicity is needed. The so-called triggering parameters could be investigated by evaluating the lag time between stress change and seismic events through ML approaches that look for more complex relationships. Since stress measurements are not available per se, it will need to be formed from a catalog of different data types such as full stimulation parameters, injection/production histories, microseismicity, and geologic context. For ML to help evaluate the contributions to triggering, extensive data from multiple data types are needed, such that an Energy Data eXchange-level dataset could help to jumpstart this type of analysis. Despite the promise of ML to help with this style of investigation, ML approach complexity needs to be evaluated versus the practicality of understanding the results. Another example where ML could help to understand physical mechanisms is by using it to improve the estimations of bottom hole pressures from surface pressures. There is very limited bottom hole pressure data right now, but if high sample rate was available at a few wells or lower sample rate at many wells in a focused area, then ML could help establish relationships between surface and bottom hole pressures that could be implemented more broadly. Improved ability to make real-time measurement of bottom hole pressures/stress, even in a few cases, could significantly improve the evaluation of the driving mechanisms of seismicity over broader areas. Recent work has proposed that large volume fluid injections can create substantial changes in subsurface pore fluid pressures several tens of kilometers away from the well, but the downhole measurements to test this idea are very limited.

Finally, a key long-term goal worth noting is to better understand to what degree the relationships between observations and induced seismicity are site-specific versus transferrable. This essentially is unknown right now, but a better understanding of this is necessary to have any chance at characterizing a site prior to drilling a well or evaluating a well that has begun operations but has yet to induce significant levels of seismicity. ML is likely to be essential for testing what is site-specific versus what is transferable once the datasets in multiple regions reach critical mass for making comparisons.

Case Study 1: ML to Enhancing Real-Time Evaluation of Induced Seismicity

Traditional earthquake catalogs employ a network association of detections from elevated short-term amplitude relative to long-term amplitude. While this is generally applicable to a wide array of source types and the processing is computationally efficient, the approach has difficulty detecting smaller events that typically occur in the early stages of injection induced seismicity. Matched-filter (template matching) approaches have been used to improve sensitivity to the smallest events and it can be computationally efficient for scanning large datasets if a modest number of templates are utilized, but it lacks general applicability, as it only finds events similar to the chosen templates. Autocorrelation approaches can achieve both high sensitivity and general applicability, but they require enormous compute times when applied to datasets that extend over a year. Recent research has sought to develop strategies that expedite this process, with one example being a repeating seismic detector that employs unsupervised ML [9]. Agglomerative clustering of all observed signals of interest in both the frequency and time domain can expedite the process of looking for repetitive seismicity. This approach results in much improved characterizations of frequency magnitude distributions (FMDs) [35], which revealed new patterns for different geological source zones that imply differences in fault maturity and hazards depending on which faults get activated by industrial operations (Figure 10). Moving forward, these improved characterizations of induced sequences could be implemented in real-time considering how ML has cut the processing speed for years of data to less than an hour. More importantly, such rapid detection of small magnitude seismicity could enable probabilistic earthquake magnitude forecasts using frameworks such as the seismogenic index formulation.

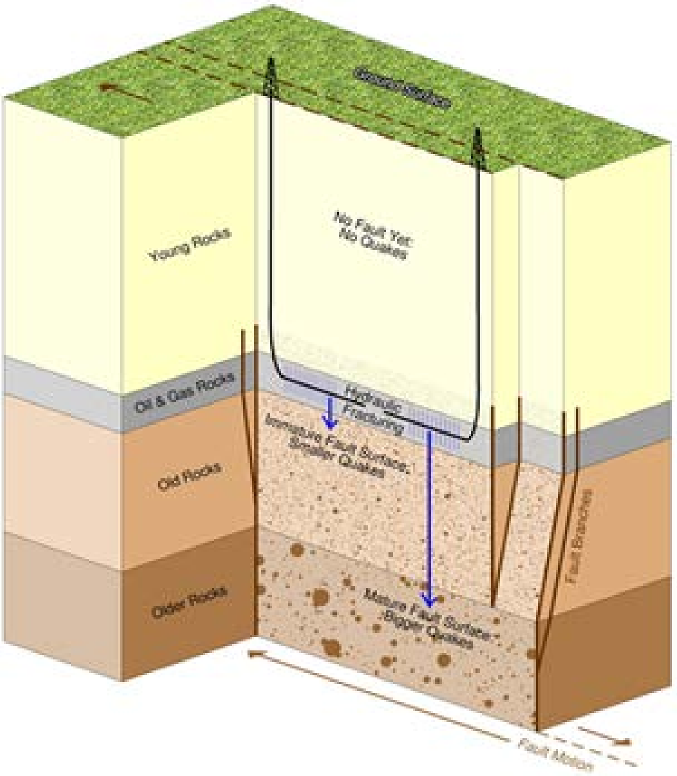

Figure 10. Schematic cross-section showing fault surfaces of a strike slip flower structure. Hydraulic fracturing can induce different seismicity in shallower sedimentary rocks compared to deeper basement rocks. Brown circles show seismic rupture areas based on the average FMDs. Blue arrows show that some hydraulic fracture wells (black curves) disturb the younger, immature part of the fault and some wells disturb the older, mature part of the fault.

Case Study 2: ML Predicts Laboratory Earthquakes

Enhanced detection of small seismic signals through increased instrumentation and improved processing provides a new opportunity to investigate whether large earthquakes can be anticipated. In a recent study, ML was applied to datasets from shear laboratory experiments, with the goal of identifying hidden signals that might precede earthquakes [13]. This study found that by listening to the acoustic signal emitted by a laboratory fault, ML can predict the time remaining before the fault fails with great accuracy (Figure 11). These predictions were based solely on the instantaneous physical characteristics of the acoustical signal and did not make use of its history. In fact, when the data points were scrambled in time, the predictions remained the same. ML identified a signal emitted from the fault zone previously thought to be low-amplitude noise, which, in turn, enabled failure forecasting throughout the laboratory earthquake cycle. The researchers inferred that this signal originates from continuous grain motions of the fault gouge as the fault blocks displace [10, 12]. Applying ML approaches to continuous seismic data such as this may lead to significant advances in identifying currently unknown signals. This holds particular promise in injection induced seismicity where swarms of small magnitude seismicity have already been observed, but observations from tectonic environments such as subduction zones have revealed more continuously active fault tremor [36], which may be analogous to the laboratory acoustic signals. Research efforts utilizing ML to better characterize continuous seismic recordings hold the potential to provide new insights into fault physics and in placing bounds on fault failure times.

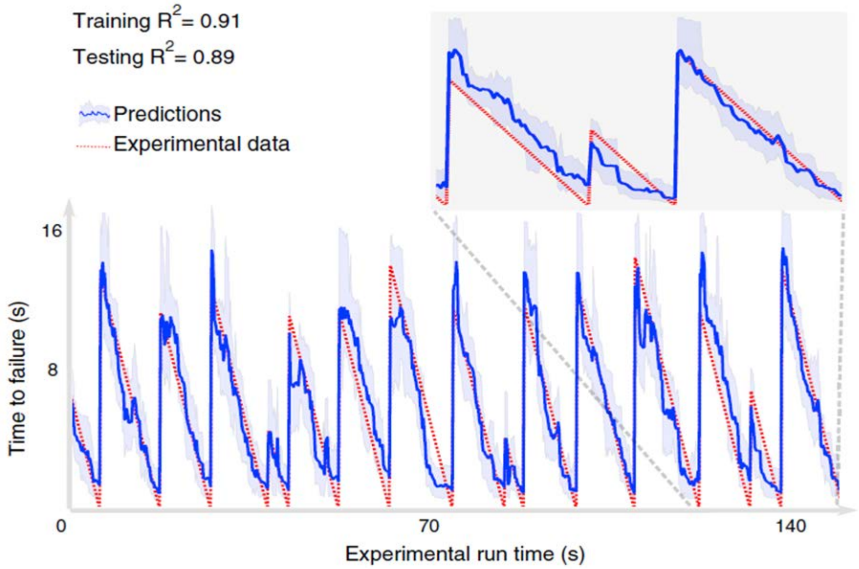

Figure 11. Forecasting quakes in the laboratory from a bi-axial shear device. The figure shows the time remaining before the next failure predicted by ML [13]. The red lines show the actual time before failure (y-axis) versus experimental run time (x-axis). The red dashed line shows the time remaining before the next failure (derived from the shear stress data), and the blue line shows the corresponding prediction of the random forest regression model (derived exclusively from the “instantaneous” acoustic data; each blue point is a “now” prediction independent of past or future observations). The blue-shaded region indicates the 5th and 95th percentiles of the forecast. The inset emphasizes that ML does well forecasting slip times even with aperiodic data. The ML trained on ≈150 s of data (≈10 slip events), and tested on the following ≈150 s, shown here. Very recently, similar forecasting of failures have been made in Earth of slow slip in Cascadia [9], and in injection experiments in southern France [37].

Appendix 4. Expanded Content for Enabling Tools and Approaches

As discussed in Enabling Tools and Approaches, supervised learning can be further subdivided into: (1) regression problems, where the response variable is continuous, or (2) classification problems, where the response variable is categorical. For both cases, predictor variables may be continuous and/or categorical [14]. Mishra and Lin outlined an example relevant to subsurface applications, where building a predictive model for the cumulative annual production in the first 12 months would be considered a regression problem, whereas determining the factors responsible for separating the top 25% of the wells ("good") from the bottom 25% ("bad") in terms of cumulative production is a classification problem [16]. Conversely, unsupervised learning algorithms draw inferences from non-labeled data (absent to known or labeled outcomes). These algorithms are geared toward discovering unknown data patterns and data structures [32]. Reinforcement learning is based on an "agent" rather than a classic model [32], learning and gaining insight from an environment through interactions, and receiving rewards (positive or negative) for performing different types of actions [38] (Figure 6). The appropriate algorithm(s) ultimately selected will be based on several factors. The schematic shown in Figure 6 (under Enabling Tools and Approaches) provides insight into selection of the most appropriate algorithm based on problem definition and data available/expected. Table 3 below provides insight into the more commonly used data-driven modeling techniques (focused on both regression and classification problems) for the petroleum geosciences [14, 16, 22].

Algorithm	Description
Linear Regression	Ordinary least-squares multiple linear regression
Classification and Regression Trees	Binary decision trees where the predictor space is split into nested rectangular regions, each with a constant value or categorical label for the response
Artificial Neural Network	Connecting inputs through layers of connected nodes to generate final outputs, like the human brain and nervous system
Random Forest	Collective of simple regression trees, each of which is trained using a random subset of observations and predictors
Bayesian Ridge	Linear regression using regularization factors that are generated by probabilistic models
Lasso	Linear regression with regularization and variable selection
Elastic Net	Combines approaches used in the lasso and ridge regression methods
Self-Organizing Map	Classifies data in such a way so the multi-dimensional input dataset is converted to a lower-dimensional grid space without losing geometrical relationship among data points
Support Vector Machine (SVM)	Transmutes data into another space in which a linear regression or linear classification-style approaches can then be used to model the data
Gradient Boosting Machine (GBM)	Ensemble of regression trees which are trained sequentially, with each new tree designed to address shortcomings in predictions made by earlier trees

Table 3. Common predictive modeling techniques for regression and classification problems that could be relevant in subsurface applications [22].

Working through a decision framework (like the one presented in Figure 12) based on the unique considerations of a given project (i.e., objective and data types and volume) should provide insight into more promising algorithms to consider for selection moving forward. However, more than one decision branch may apply given the specific project, and in other instances, none of the branches will align perfectly [39]. Figure 12 below provides insight into the more commonly used data-driven modeling techniques (focused on both regression and classification problems) for the petroleum geosciences [14, 16, 22].

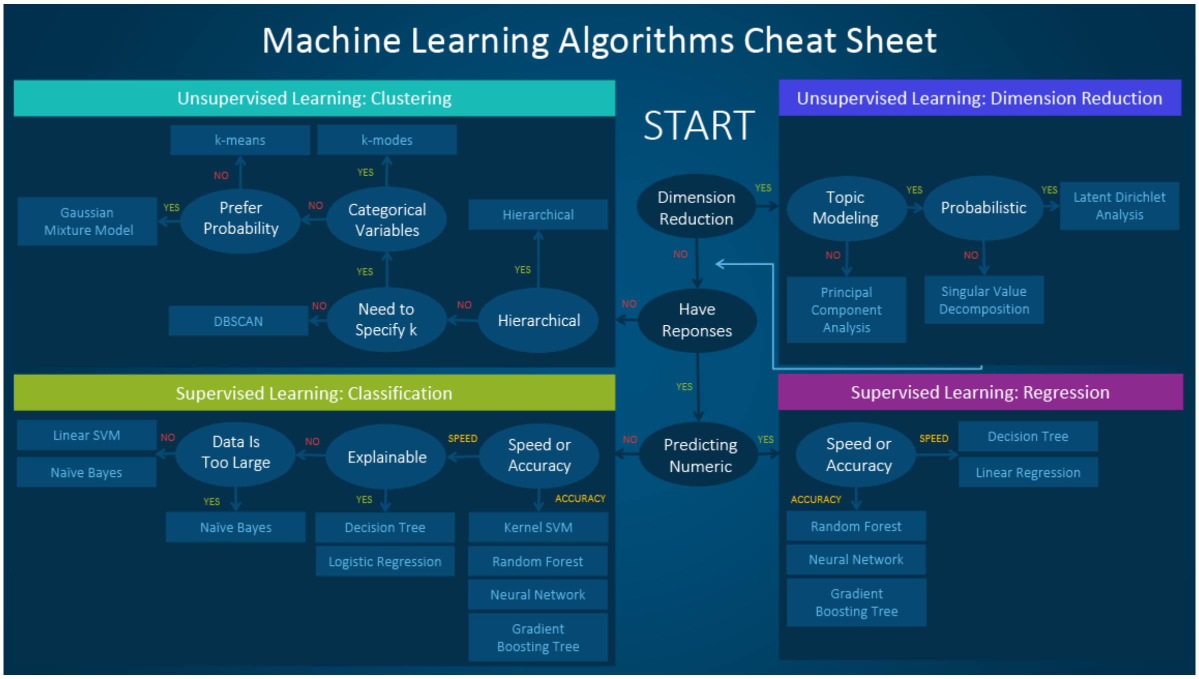

Figure 12. Algorithm selection decision flow diagram concept [29, 39]

Model Evaluation and Validation

Three common metrics for evaluating the quality-of-fit, per Mishra and Lin, include [16] (1) average absolute error (AAE), (2) mean squared error (MSE), and (3) pseudo-R^2 [40]. The AAE is defined as the average magnitude of the difference between the true response and predicted response. MSE measures the average squared difference between observations and their corresponding predictions, rather than the absolute value like AAE. A common variant of MSE is the root mean square error (RMSE), which is simply the square root of MSE. The third metric, pseudo-R^2, compares the sum of squared differences between the true responses and predicted responses to the overall sum of squares.

A common approach to evaluating the quality-of-fit for a model is to generate a scatterplot for the actual responses or observations from testing or independent datasets, as compared to the corresponding predicted responses. If the points in the scatterplot lie close to the 1:1 line (at 45°), this indicates a good model fit to the training data [14]. This type of validation step needs performed against the testing or independently-derived datasets, which were not used to train the model, and would better represent model accuracy against unbiased data and help indicate if any overfitting had occurred [16].

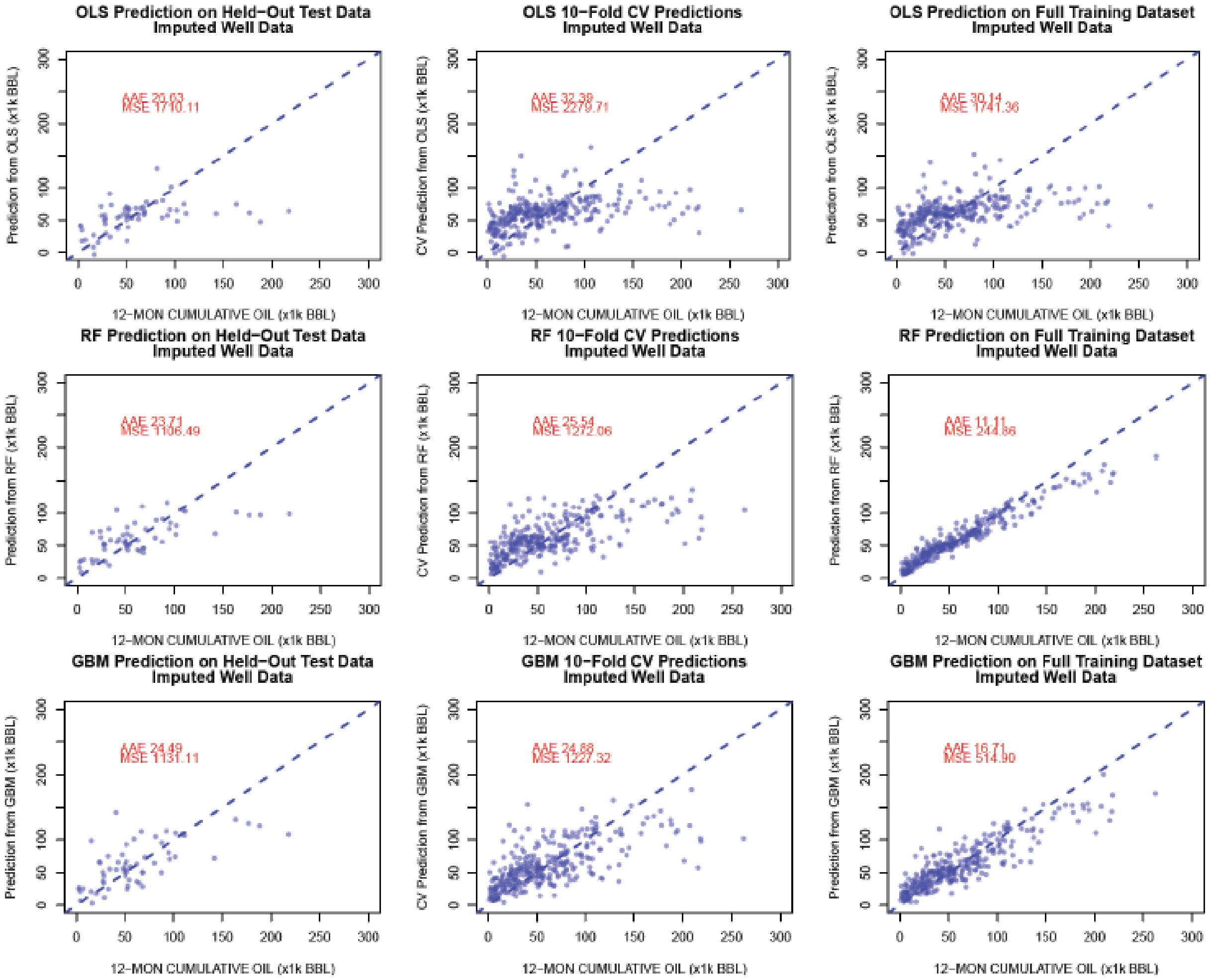

Figure 13. Comparison of the performance of different models using AAE and MSE built in various algorithm approaches. A subset of data presented in Schuetter et al., 2015 [21].

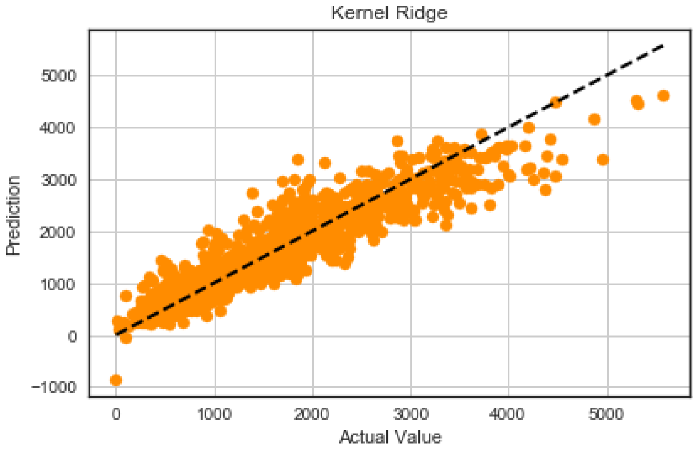

Figure 14. Example quality-of-fit of the kernel ridge model comparing predictions to actual well productivity (measured as first-year production data in million cubic feet equivalent) from Shih et al., 2018 [22]. In this specific example, an R^2 score against the training dataset was 0.85 for the kernel ridge model.

If the developed model underperforms, it might be necessary to utilize more advanced strategies to improve the performance of the model, utilize a different algorithm(s), and possibly supplement datasets with additional data (either in quantity or complexity). Further improvement can possibly be accomplished by tuning certain model parameters. Automated processes that rely on cross-validation have been suggested in the literature as an approach to quantify the impact of parameter tuning [16, 14, 41]. Examples are provided in Shih et al. [22]. In this specific example, an R2 score against the training dataset was 0.85 for the kernel ridge model.

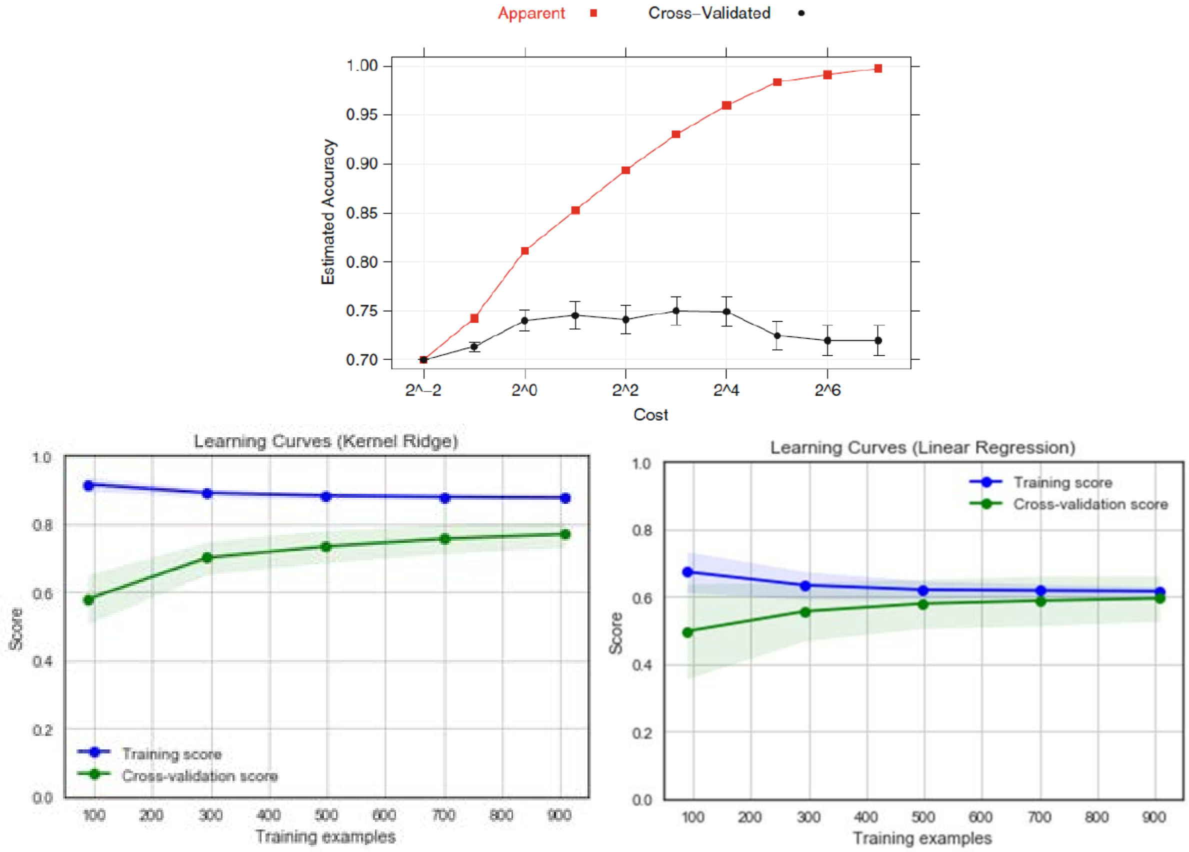

Figure 15. Examples from cross-validated ML parameter tuning: (top) automatic tuning of SVM model parameters using cross-validated RMSE [14]; (bottom) learning curves to evaluate predictive model quality-of-fit convergence with cross-validation against training sample sizes from two different regression algorithms [22].

For models that are more complex, initial conditions play a significant role in the determination of the outcome of training. Differences can be seen depending on whether a model starts off training with values initialized to zeroes versus some distribution of values, which then leads to the question of which distribution is to be used. Since there are many considerations at this phase of training, it is important that what makes a model good is defined. These parameters are referred to as hyperparameters. The adjustment or tuning of these parameters depends on the dataset, model, and the training process.

Appendix 5. Workshop Attendees

First Name	Last Name	Organization
Seth	**King**	AECOM
Michael	**Prinkey**	Aeolus Technologies
Tim	**Kneafsey**	Lawrence Berkley National Laboratory (LBNL)
Martin	**Schoenball**	Lawrence Berkley National Laboratory (LBNL)
Christopher	**Sherman**	Lawrence Livermore National Laboratory (LLNL)
George	**Guthrie**	Los Alamos National Laboratory (LANL)
Paul	**Johnson**	Los Alamos National Laboratory (LANL)
Velimir	**Vesselinov**	Los Alamos National Laboratory (LANL)
Vic	**Baker**	MATRIC
Grant	**Bromhal**	National Energy Technology Laboratory (NETL)
Jared	**Ciferno**	National Energy Technology Laboratory (NETL)
Dustin	**Crandall**	National Energy Technology Laboratory (NETL)
Robert	**Dilmore**	National Energy Technology Laboratory (NETL)
Antonio	**Ferreria**	National Energy Technology Laboratory (NETL)
Don	**Remson**	National Energy Technology Laboratory (NETL)
Traci	**Rodosta**	National Energy Technology Laboratory (NETL)
Kelly	**Rose**	National Energy Technology Laboratory (NETL)
Madhava	**Syamlal**	National Energy Technology Laboratory (NETL)
Hector	**Santos**	Oak Ridge National Laboratory (ORNL)
Casie	**Davidson**	Pacific Northwest National Laboratory (PNNL)
Tim	**Johnson**	Pacific Northwest National Laboratory (PNNL)
Alexandre	**Tartakovsky**	Pacific Northwest National Laboratory (PNNL)
Giorgia	**Bettin**	Sandia National Laboratory (SNL)
Lisa	**Linville**	Sandia National Laboratory (SNL)
Alan	**Cohen**	U.S. DOE Office of Fossil Energy
Darin	**Damiani**	U.S. DOE Office of Fossil Energy
Steven	**Lee**	U.S. DOE Office of Science
Michael	**Matuszewski**	AristoSys, LLC
Elizabeth	**Eide**	American Geosciences Institute, The National Academies of Sciences, Engineering, and Medicine
Kaushik	**Dayal**	Carnegie Mellon University
Andrew	**Gellman**	Carnegie Mellon University

First Name	Last Name	Organization
Matteo	Pozzi	Carnegie Mellon University
Mitchell	Small	Carnegie Mellon University
Anna	Siefken	Carnegie Mellon University
Erik	Ydstie	Carnegie Mellon University
Jared	Smith	Cornell University
Bradford	Hager	Massachusetts Institute of Technology
Mike	Brudzinski	Miami University
Miranda	Joyce	Rice University
Parthib	Rao	Rice University
Alireza	Shahkarami	Saint Francis University
Charles	Gorecki	University North Dakota, Energy & Environmental Research Center (UND EERC)
Jose	Torres	University North Dakota, Energy & Environmental Research Center (UND EERC)
Masoud	Kalantari	University of Kansas and ICPRs LLC
Andrew	Bunger	University of Pittsburgh
Bill	Harbert	University of Pittsburgh
Fred	Aminzadeh	University of Southern California
Seyyed	Hosseini	University of Texas/Bureau of Economic Geology
Neeraj	Gupta	Battelle
Srikanta	Mishra	Battelle
Leon	Wu	BriskPoint Inc.
Yan	Pan	Chevron
Simon	Shaw	ConocoPhillips
Lewis	Matthews	CrownQuest Operating, LLC
Jalal	Jalali	EQT
Gavin	Liu	GE Baker Hughes
Sallie	Greenberg	Illinois State Geological Survey
Andrea	Cortis	Pioneer Natural Resources
Carl	Carlson	Range Resources
Joe	Frantz	Range Resources
Ryan	Tyree	Range Resources
Brian	Parsonnet	Seeq Corporation

Appendix 6. Workshop Schedule

AGENDA

Real-Time Decision-Making for the Subsurface Workshop

Wilton E. Scott Institute for Energy Innovation at Carnegie Mellon University

Pittsburgh, PA

July 17–July 18, 2018

TUESDAY, JULY 17

Plenary Session

7:30–8:15 a.m. **Registration and Continental Breakfast** (Singleton Room in Roberts Hall)

8:15–9:00 a.m. **Welcome Remarks**

Emcee: **Andy Gellman**, Lord Professor of Chemical Engineering, co-Director, Wilton E. Scott Institute for Energy Innovation, Carnegie Mellon University

Speaker: **Steve Winberg**, Assistant Secretary for Fossil Energy, DOE

Speaker: **Randall Gentry**, Deputy Director and Chief Technology Officer, DOE NETL

Speaker: **Ramayya Krishnan***,* Dean of Heinz College of Information Systems and Public Policy, Carnegie Mellon University

Speaker: **Jay Whitacre*,*** Director, The Wilton E. Scott Institute for Energy Innovation, Carnegie Mellon University

9:00–10:00 a.m. **Primer Presentations**

Three (3) approximately 15-minute presentations

1. **Resource Recovery and Reservoir Management**
   Speaker: **Grant Bromhal**, Senior Fellow, NETL
2. **Autonomous Monitoring Systems**
   Speaker: **Charles Gorecki**, Director of Subsurface R&D, University of North Dakota EERC
3. **Induced Seismicity**
   Speaker: **Paul Johnson**, Senior Fellow, LANL

10:00–10:15 a.m. **Break**

10:15–11:30 a.m. **Technical Plenary Talks/Industry Case Studies**

Four to five (4–5) approximately 15–20-minute presentations

Industry examples of approaches currently being used in the field, case studies, success stories, future plans, and address-related questions

Plenary: ML and Big Data Approaches

Speaker: **Srikanta Mishra**, Senior Research Leader, Battelle

Case studies (e. Data Sharing; Missing Data; Successful Techniques, Others TBD)

Speaker 1: **Fred Aminzadeh**, Research Professor,
University of Southern California

Speaker 2: **Ryan Tyree**, Production Engineering Manager,
Range Resources

Speaker 3: **Andrea Cortis,** Innovation Data Science Manager,
Pioneer Natural Resources

Speaker 4: **Mike Brudzinski,** Geology Post-Doctoral Fellow,
Miami University

11:30–11:45 a.m. **Charge to Breakout Groups and Split into Groups:**

1. **Resource Recovery and Reservoir Management**

   Lead: **Grant Bromhal**, Senior Fellow, NETL

   Co-lead: **Lewis Matthews,** Data Scientist,
   CrownQuest Operating, LLC

2. **Autonomous Monitoring Systems**

   Lead: **George Guthrie**, Focus Area Lead for Geosciences, LANL

   Co-lead: **Alan Cohen**, Director, Office of Oil and Natural Gas Research, DOE FE

3. **Induced Seismicity**

   Lead: **Paul Johnson,** Senior Fellow, LANL

   Co-lead: **Mike Brudzinski**, Geology Post-Doctoral Fellow,
   Miami University

Breakout Sessions

11:45 a.m.–12 p.m.	**Walk to Breakout Rooms** (Scott Hall)
12:00–2:00 p.m.	**First Breakout Session: TARGETS** (working boxed lunch)
2:00–2:15 p.m.	**Break**
2:15–3:45 p.m.	**Second Breakout Session: DATA AVAILABILITY**
3:45–4:00 p.m.	**Break/Return Walk to Plenary Room** (Singleton Room in Roberts Hall)

Day One Closing Discussion

4:00–5:00 p.m.	**Report Outs from Breakout Sessions**
5:00–7:00 p.m.	**Reception** (Singleton Room in Roberts Hall and adjacent foyer)

WEDNESDAY, JULY 18

Plenary Session

7:30–8:00 a.m.	**Continental Breakfast** (Singleton Room in Roberts Hall)
8:00–8:45 a.m.	**Discuss Results from Day One and Charge for Day Two Breakouts**

Breakout Sessions

8:45–9:00 a.m.	**Walk to Breakout Group Rooms** (Scott Hall)
9:00–11:15 a.m.	**Third Breakout Session: BARRIERS & MISSING DATA**
11:15 a.m.–12 p.m.	**Break**
12:00–2:00 p.m.	**Fourth Breakout Session: TOOLS & TECHNIQUES** (working boxed lunch)
2:00–2:30 p.m.	**Wrap-up Discussion in Breakout Sessions to Prepare Report Outs**
2:30–3:00 p.m.	**Break/Return Walk to Plenary Room** (Singleton Room in Roberts Hall)

Day Two Closing Discussion

3:00–4:30 p.m.	**Report Outs from Breakout Sessions** *30 minutes for each working group*
4:30–5:00 p.m.	**Closing Discussion**
5:00 p.m.	**Adjourn**