Introduction to this special section: Data analytics and machine learning

Mike Davidson[1]

The explosion of data types and volumes we are experiencing in the oil field, particularly in unconventional reservoir development, is stretching the capabilities of traditional manual workflows. The complex interactions between dynamic reservoir properties and the many-faceted well completions process are governed by complex physics that may be only partially understood. Furthermore, our best attempts to perform controlled experiments can often produce highly varied results, suggesting the stochastic nature of production from unconventionals and that we may benefit from adding new tools to our toolkit of deterministic, physics-based analysis.

Other industries face similar challenges, from Internet companies searching to understand subtle consumer preferences that will drive incremental sales to pharmaceutical companies looking to identify characteristics of patients that may respond to a new drug treatment. Many techniques used in these other industries can be directly utilized to understand oil field data and to optimize cost of supply; however, our industry has been relatively slow in adopting these new approaches.

The methods we discuss in this special section go by several names, such as analytics, data mining, data science, multivariate analysis, or machine learning — all are nearly synonymous, and there are others. Particular methods range from traditional statistical analysis to new algorithms including gradient boosted regression trees, text mining, and deep learning. These methods allow us to make sense of lots of data with many variables while avoiding the biases that humans can bring to any analysis where data set sizes may be reduced to just examining a few points or a few variables. At the heart of these methods are best practices in which hypotheses are formulated from the data, tested, and a sense of prediction accuracy is developed to inform business users whether they can proceed with the analysis result.

Overarching the opportunities to leverage data from many disciplines is the inexorable march to automate tasks when a machine's performance can be better than, or close to, human performance. This phenomenon is seen clearly in manufacturing industries but is also fully expected to affect knowledge workers in the coming decades. In the near term, these technologies are expected to enhance the productivity of subject matter experts by performing the majority of time-consuming analysis and only notifying a human when additional input is needed. In the long-term, who knows? We look forward to pushing the barriers to understand what we can automate.

To illustrate the potential of machine learning for seismic imaging — a domain many readers will be familiar with — Araya-Polo et al. show how the entire imaging workflow could be bypassed and faults detected directly from field data. Perozzi et al. demonstrate the utility of the analytics approach in gold mining resource estimation and contrast the effectiveness of a number of different commonly employed algorithms.

Ruths et al. discuss a workflow to integrate petrophysical and mechanical properties along horizontal wellbores to understand production differences and to suggest potential for optimizing cost. To address what is sometimes observed to be the "black-box" nature of machine learning algorithms, Alumbaugh and Schnetzler apply what may be a more interpretable algorithm to integrate potential fields and well logs in an exploration setting.

Using synthetic data sets, Cao and Roy offer a glimpse into a workflow that may unify empirical analysis with our foundational reservoir physics by closing a loop between observed 4D seismic responses, simulated reservoir property changes, and predictive analytics. Sánchez Galvis et al. demonstrate how seismic field data might be intelligently partitioned for further processing or analysis using a *k*-means clustering algorithm.

Many of the machine learning advances in the past decade or so have been driven by enabling technologies such as smarter algorithms, faster CPUs and GPUs, larger data sets to "train" the machines, along with the ability to scale the analysis to any size. Huang et al. describe one such enabling technology and its application for fault detection.

A considerable challenge faced by many working geoscientists is the time spent searching for, wrangling with, and making sense of large legacy databases, many of which may lack a natural organizational structure. This unstructured data may include such information as field notes from ops-geologists on vintage mud-logs — potentially valuable if automated methods could be employed to make sense of it. Blinston and Blondelle show how machine learning could be employed in the management of databases.

Finally, Hall provides some thought-provoking use cases and points out that machine learning is a rapidly developing field. Our industry, in particular, is on the steep part of the learning curve. Now is the time to invest in understanding the potential, the pitfalls, and the limitations of machine learning.

The contributions in this special section clearly demonstrate the tremendous potential and opportunities that lie ahead for adopting machine learning techniques and further developing workflows for the purpose of solving subsurface problems. In an era in which the value of multidiscipline (multiphysics) and data integration have been widely recognized, but the mechanism or the engine to accomplish such integration is still being defined, the value of machine learning applications is far reaching and ranges from solving complex physical problems to improved efficiency in multiple dimensions, directly impacting our bottom line. TLE

[1]ConocoPhillips.

http://dx.doi.org/10.1190/tle36030206.1.