NATURAL GAS *brief*

SEPTEMBER 2018

Stanford | Natural Gas Initiative
School of Earth, Energy & Environmental Sciences and Precourt Institute for Energy

Data science, artificial intelligence, and related emerging technologies have made, and have the potential to make, major impacts across the natural gas value chain. Transferring rapidly developing technologies to industry-specific application requires a strong blend of data science, artificial intelligence, and subject matter expertise.

A personal journey through data science—

Demystifying data science and what it means for engineering our natural resources

By Jef Caers, Stanford University

I have recently come to understand that I am a data scientist; who knew? My own journey started with a PhD dissertation on the valuation of diamond deposits: How can we estimate and quantify uncertainty on the total amount of stones, their sizes, and dollar values in the entire deposit, given a limited number of samples (~1000 stones) taken from it. In Earth resources, we have both a small data and a big data problem. Appraising relies on a few samples to estimate a large population (the resource), or we may have big data (e.g., geophysics, continuous monitoring) that may, however, carry varying degrees of information about what we need to know (movement of fluids, volume of mineable ore). It is in understanding this dichotomy that we can understand the role of data science (statistical science and machine learning) in the engineering of natural resources. ▸

ABOUT THE AUTHOR

Jef Caers

Jef Caers received both an MSc ('93) in mining engineering / geophysics and a PhD ('97) in engineering from the Katholieke Universiteit Leuven, Belgium. Currently, he is Professor of Geological Sciences (since 2015) and previously Professor of Energy Resources Engineering at Stanford University, California, USA. He is also director of the Stanford Center for Earth Resources Forecasting, an industrial affiliates program in decision making under uncertainty with ~20 partners from the Energy Industry. Dr. Caers' research interests are quantifying uncertainty and risk in the exploration and exploitation of Earth Resources. Jef Caers has published in a diverse range of journals covering Mathematics, Statistics, Geological Sciences, Geophysics, Engineering and Computer Science. He was awarded the Vistelius award by the IAMG in 2001, was Editor-in-Chief of Computers and Geosciences (2010-2015). Dr. Caers has received several best paper awards and written four books entitled *Petroleum Geostatistics* (SPE, 2005) *Modeling Uncertainty in the Earth Sciences* (Wiley-Blackwell, 2011), *Multiple-point Geostatistics: stochastic modeling with training images* (Wiley-Blackwell, 2015) and *Quantifying Uncertainty in Subsurface Systems* (Wiley-Blackwell, 2018). Dr. Caers was awarded the 2014 Krumbein Medal of the IAMG for his career achievement.

For more information

Jef Caers:
https://profiles.stanford.edu/jef-caers

Natural Gas Briefs:
ngi.stanford.edu/briefs

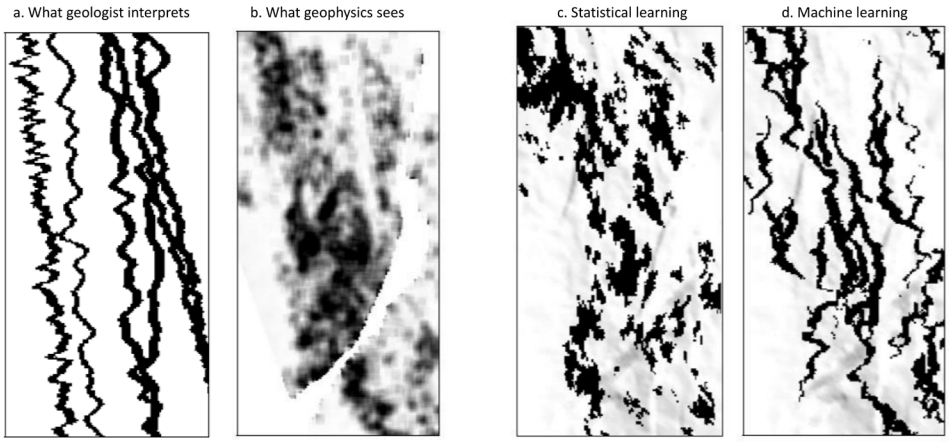
Clearly, small data problems require some form of extrapolation. This is common in many areas of the Earth sciences such as the prediction of return frequency of floods, earthquakes, eruptions, tsunamis, and… valuable diamonds. Statistical science offers such extrapolation within a rigorous mathematical framework: extreme value statistics. Extremes are what interests us most; but, what are the statistical variations of yet unseen large events? How can we infer this from limited data? While statistical science offers principled approaches to estimate these return levels, a problem occurs: physical plausibility. The statistical variation seen in Earth resources data results from the processes that created the resource. In the diamond case, stones in kimberlite deposits have a fractal distribution while stones in a river deposit have a lognormal distribution. Geological processes of mineral growth and river sorting can be used to explain these differing probability distributions. Thereby a link is made between physical processes and statistical fluctuations. Understanding this causal link is what will offer prediction accuracy.

The limitation of statistical science arises when dealing with complex high-dimensional problems; basically, problems that involve a lot of variables and uncertainties. An example would be a 3D geological model of a resource. The mathematical rigor of statistical science now gets in the way of defining flexible prediction models that can handle such complex cases. Consider an example in Figure 1, modeling of a deep-water reservoir system off the coast of West Africa. A geologist has interpreted there to be channel formation and provides a conceptual depiction (Figure 1a). Geophysical data (Figure 1b) informs (vaguely) the presence of channels. Figure 1c shows a model generated by a traditional statistical method, and Figure 1d a model generated using machine learning. What's happening?

In statistical science, probability distributions are seen as data-generating models. The statistician sees the data as being generated from abstract mathematical models, despite the fact that in reality these models do not exist, only the physical processes that generated the data. Machine learning does not use the data-generating (stochastic) model, it simply uses the data. In the situation above, a sedimentologist provides the computer scientists with a training example of a channel formation (the "data"). The goal of the computer scientists is to generate 3D geological models that look like the training example (and possibly are constrained to reservoir data such as geophysics). Powerful machine learning algorithms allow replication of what is trained and produce "realistic" patterns simply because the training set is deemed realistic. The probability model is circumvented altogether. The downside? The mathematical rigor is gone, and hence it is unclear what the generalizations of these methods are, what the effect is of ad-hoc tuning parameters, and to what extent artifacts (biases) are being created. Synopsis: Statistical science offers a rigorous mathematical framework ▶

Figure 1

a. Probability of sand occurrence as measured by geophysical data, b. conceptual geological description of channel formations in a turbidite setting, c. reservoir model generated using a traditional statistical method, d. reservoir model generated by machine learning on the conceptual model.



for inference (unbiasedness, understanding variable importance, uncertainty quantification), machine learning offers algorithm-driven computer science methods mostly for data-driven prediction (regression and classification; supervised and unsupervised).

FROM DATA TO DECISIONS

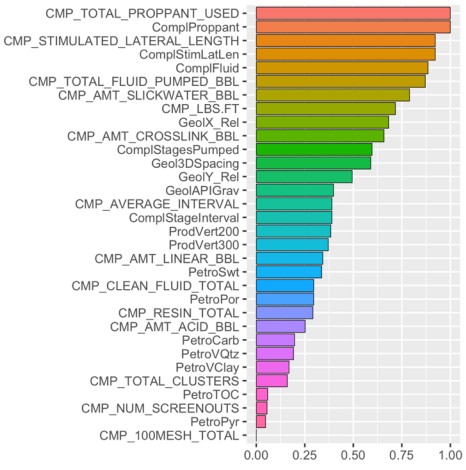
Even before the recent information technology explosion, the Earth resources (water, minerals, energy) industries have used big data (e.g., terabytes of seismic data), built billion-cell complex subsurface 3D geological models, and simulated physical processes of heat, flow, geomechanics, and chemical reactions to understand the impact of various sources of uncertainty on decision making for engineering facilities. In the upstream world, data to achieve this comes from a large variety of sources and disciplines: geology, geochemistry, geophysics, engineering, computer science, data science and decision science. In the mid-to-downstream setting, data analytics are increasingly used to optimize complex processing facilities, both for mining and oil/gas. In my own personal experience in working with these industries I have observed a siloed approach to data streams and decision making. The decision problem is often divided and conquered by partitioning into domain expertise: geology, geophysics, engineering, and management. This leads to inefficiencies and poor decision making due to the use of deterministic modeling and bottlenecks created between domains. Furthermore, modeling time often exceeds the decision timeframe. The recent advent of powerful computing and data science approaches offer a new view on the same problem. In this new paradigm, two information streams exist: 1) top down: a decision stream, making modeling purpose-driven; 2) bottom up: an uncertainty stream, rendering decision making optimal. Decision making under uncertainty requires all domain expertise and data to be handled jointly, rather than treated under a siloed approach. Monte Carlo methods (statistics science) in combination with high-performance computing infrastructure can be used to jointly consider all uncertainties. Machine learning can be used to learn from the Monte Carlo results, handle complex, high-dimensional uncertainty, discover important patterns relating data to prediction variables, and ultimately reduce uncertainty on key decision variables in a matter of days/weeks, rather than months.

EXAMPLE: SHALE RESOURCES

Data science holds many promises for appraising and developing shale resources. Many companies develop black-box approaches, often based on machine learning, to predict important decision variables such as estimated ultimate recovery (EUR). Such black-box approaches, while successful in selected cases, have limitations in terms of wider deployment; identifying overfitting is a challenge and tuning remains ad-hoc. To understand this, consider some results from an unidentified prominent shale resource in the USA. From databases of ▶

Figure 2

Sensitivity analysis based on 200 production wells in a prominent U.S. shale resource.



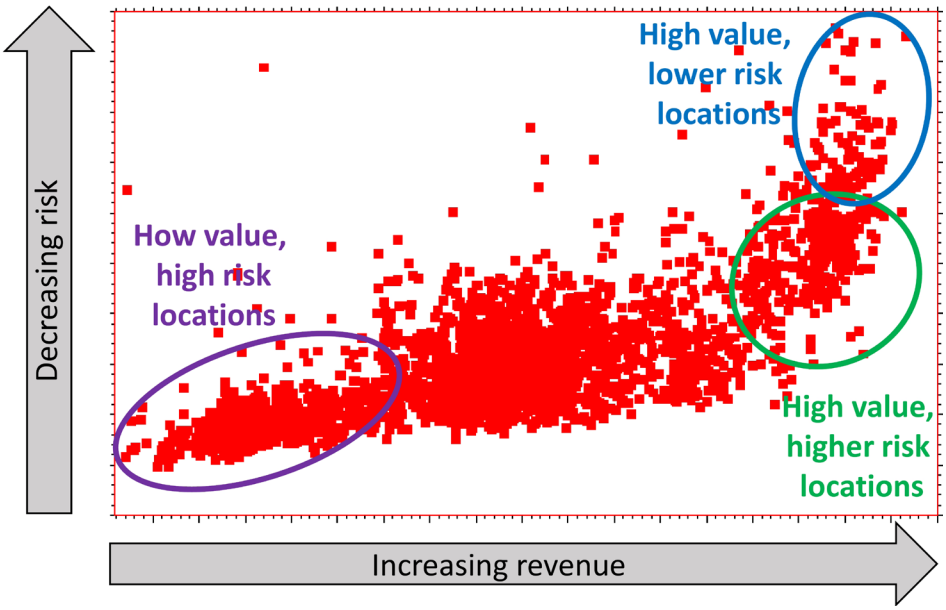
Number of Completion Stages	CMP STAGES STIMULATED
Total Amount of Injected Fluid	CMP TOTAL FLUID PUMPED GAL
Total Amount of Injected Proppant	CMP TOTAL PROPPANT USED
Stimulated Lateral Length	CMP STIMULATED LATERAL LENGTH
Total Amount of Slick Water	CMP AMT SLICKWATER BBL
Total Amount of Injected x-link fluid	CMP AMT CROSSLINK BBL
Completion Stage Interval	CompStageInterval
Total amount of linear fluid	CMP AMT LINEAR BBL
X location	GeolX Rel
Y location	GeolY Rel
Z location (depth)	GeolZ
Oil API Gravity	GeolAPIGrav
Total organic content (TOC)	PetroTOC
Clay Content	PetroVClay
Water Saturation	PetroSwt
Porosity	PetroPor
Total Amount of Quartz	PetroVQtz
Amount of Pyrite	PetroPyr

production decline, geological, and completion parameters, one can run a sensitivity analysis to understand what is impacting production. Figure 2 shows such a typical result. It appears the production decline is dominated by completion parameters, while geological parameters rank at the bottom. Does geology not matter? Not really: The geological parameters in these databases are often averages, or values picked from a smooth map. Additionally, many important geological factors such as fracture density are rarely measured in shale systems, simply because of cost. Consequently, the system may therefore be over-engineered because of lack of measurements and the understanding generated from them.

Data-based sensitivity analysis may also provide important constraints on the effect of engineering choices, such as the amount of sand pumped per foot, the number of stages, choke size, and well spacing on production decline. To further understand this and make prediction of optimal drilling locations and strategies of completion to use at those locations, a principled "white box" approach combining statistical science and machine learning can be employed. As alluded to in the first section, machine learning is best used to discover relationships in high-dimensional problems; here, high dimensions simply refer to the large amount of geological and completion parameters. However, to make predictions and quantify uncertainty, a principled statistical approach is needed. The shale resource prediction problem is essentially a multivariate space-time statistical problem: Multiple phases (oil, water, gas) decline over time with drilling in geographical space. To account properly for all of these relationships, one can combine functional data analysis (time) with geostatistics (space) to make predictions of EUR at yet undrilled locations and quantify uncertainty. Uncertainty on these (statistical) predictions is improved by including machine learning that relates geological and completion factors to production decline. The results of this combined statistics–machine learning approach can then be summarized in a so-called efficient frontier plot (risk vs. return). Such plots are also used in analyzing rock portfolios to assess risk (uncertainty future price) vs. return (median stock price). Figure 3 shows such as plot, where uncertainty (risk, y-axis) is quantified using statistical methods; prediction (return, x-axis) is mostly dependent on machine learning.

Figure 3

Risk vs. return estimated for undrilled locations in a prominent U.S. shale resource, allowing the identification and ranking of locations that includes uncertainty quantified using statistical science (units removed because of confidentiality).



LIMITATIONS OF DATA SCIENCE

First, in appraising new resources no or very few production wells or pilot tests have been drilled, hence the use of data analytics to understand risk vs. return becomes questionable. The data science question now is ▶

> *"In Earth resources, we have both a small data and a big data problem."*

more fundamentally geological: To what extent can previous basins be used as analogies for a new basin? Now, an additional uncertainty about similarity, measured though understanding of basin geological processes (source, maturation history), exists. Such degree of similarity is handled typically with Bayesian approaches (statistical science) assigning prior probability of similarity (weights) to data from previously drilled basins.

Second, data science approaches cannot identify what additional data should be used to improve prediction making. In decision science, this is known as the value of information problem. To address this question, one will need to understand what yet unmeasured properties impact production decline and what data can be used to reduce its uncertainty. Currently, speculations are made around acquiring geophysics (to map structural features) or borehole-oriented tools that can measure natural fracture properties. Formal modeling frameworks for establishing this value of information are yet to be established.

Third, data science approaches can at best identify correlating factors; e.g., correlation between production decline and number of stages. Causal relationships can only be fully quantified and understood though mechanistic (physical) models. The problem in shales often is that the "mechanisms" of flow in porous media are not fully understood. It seems that any future progress will therefore require developing a data scientific method that can combine Monte Carlo on improved mechanistic models with production data acquired in the actual field.

IN SUMMARY

Data science offers an enormous potential for the Earth resources industry. Such potential can be realized by looking beyond the black-box approaches that have had successes in human-oriented applications. The uniqueness of dealing with the physical world and the uncertainty associated with the lack of access and understanding will require developing well-understood white-box approaches that aim to enhance—rather than attempt to replace—existing engineering knowledge. ▲

THE NATURAL GAS INITIATIVE AT STANFORD

Major advances in natural gas production and growth of natural gas resources and infrastructure globally have fundamentally changed the energy outlook in the United States and much of the world. These changes have impacted U.S. and global energy markets, and influenced decisions about energy systems and the use of natural gas, coal, and other fuels. This natural gas revolution has led to beneficial outcomes, like falling U.S. carbon dioxide emissions as a result of coal to gas fuel switching in electrical generation, opportunities for lower-cost energy, rejuvenated manufacturing, and environmental benefits worldwide, but has also raised concerns about global energy, the world economy, and the environment.

The Natural Gas Initiative (NGI) at Stanford brings together the university's scientists, engineers, and social scientists to advance research, discussion, and understanding of natural gas. The initiative spans from the development of natural gas resources to the ultimate uses of natural gas, and includes focus on the environmental, climate, and social impacts of natural gas use and development, as well as work on energy markets, commercial structures, and policies that influence choices about natural gas.

The objective of the Stanford Natural Gas Initiative is to ensure that natural gas is developed and used in ways that are economically, environmentally, and socially optimal. In the context of Stanford's innovative and entrepreneurial culture, the initiative supports, improves, and extends the university's ongoing efforts related to energy and the environment.

Join NGI

The Stanford Natural Gas Initiative develops relationships with other organizations to ensure that the work of the university's researchers is focused on important problems and has immediate impact. Organizations that are interested in supporting the initiative and cooperating with Stanford University in this area are invited to join the corporate affiliates program of the Natural Gas Initiative or contact us to discuss other ways to become involved. More information about NGI is available at ngi.stanford.edu or by contacting the managing director of the initiative, Bradley Ritts, at ritts@stanford.edu.