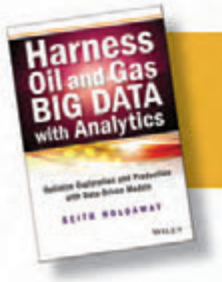
Harness Oil and Gas BIG DATA with Analytics

Optimize Exploration and Production with Data-Driven Models

KEITH HOLDAWAY

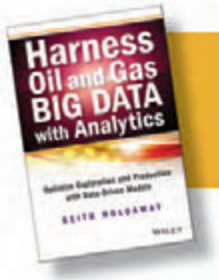
WILEY



From *Harness Oil and Gas Big Data with Analytics*. Full book available for purchase here.

Contents

From *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models* by Keith R. Holdaway. Copyright © 2014, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Harness Oil and Gas Big Data with Analytics*. Full book available for purchase here.

CHAPTER 1

Fundamentals of Soft Computing

There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy.

William Shakespeare: *Hamlet*

The oil and gas industry has witnessed a compelling argument over the past decade to adopt soft computing techniques as upstream problems become too complex to entrust siloed disciplines with deterministic and interpretation analysis methods. We find ourselves in the thick of a data avalanche across the exploration and production value chain that is transforming data-driven models from a professional curiosity into an industry imperative. At the core of the multidisciplinary analytical methodologies are data-mining techniques that provide descriptive and predictive models to complement conventional engineering analysis steeped in first principles. Advances in data aggregation, integration, quantification of uncertainties, and soft computing methods are enabling supplementary perspectives on the disparate upstream data to create more accurate reservoir models in a timelier manner. *Soft computing* is amenable, efficient, and robust as well as being less resource intensive than traditional interpretation based on mathematics, physics, and the experience of experts. We shall explore the multifaceted benefits garnered from the application of the rich array of soft computing techniques in the petroleum industry.

CURRENT LANDSCAPE IN UPSTREAM DATA ANALYSIS

What is *human-level artificial intelligence*? Precise definitions are important, but many experts reasonably respond to this question by stating that such phrases have yet to be exactly defined. Bertrand Russell remarked:

> I do not pretend to start with precise questions. I do not think you can start with anything precise. You have to achieve such precision as you can, as you go along.[1]

The assertion of knowledge garnered from raw data, which includes imparting precise definitions, invariably results from exhaustive research in a particular field such as the upstream oil and gas (O&G) disciplines. We are seeing four major trends impacting the *exploration and production* (E&P) value chain: Big Data, the cloud, social media, and mobile devices; and these drivers are steering geoscientists at varying rates toward the implementation of soft computing techniques.

The visualization of Big Data across the E&P value chain necessitates the usage of Tukey's suite of exploratory data analysis charts, maps, and graphs[2] to surface hidden patterns and relationships in a multivariate and complex upstream set of systems. We shall detail these visual techniques in Chapters 3, 4, and 9 as they are critical in the data-driven methodologies implemented in O&G.

Artificial neural networks (ANN), fuzzy logic (FL), and genetic algorithms (GA) are human-level artificial intelligence techniques currently being practiced in O&G reservoir management and simulation, production and drilling optimization, real-time drilling automation, and facility maintenance. Data-mining methodologies that underpin data-driven models are ubiquitous in many industries, and over the past few years the entrenched and anachronistic attitudes of upstream engineers in O&G are being diluted by the extant business pressures to explore and produce more hydrocarbons to address the increasing global demand for energy.

Digital oilfields of the future (DOFFs) and intelligent wells with multiple sensors and gauges are generating at high velocity a plethora of disparate data defining a complex, heterogeneous landscape such as a reservoir-well-facility integrated system. These high-dimensionality data are supplemented by unstructured data originating from social media activity, and with mobile devices proving to be valuable in field operations and cloud computing delivering heightened flexibility and increased performance in networking and data management, we are ideally positioned to marry soft computing methodologies to the traditional deterministic and interpretive approaches.

Big Data: Definition

The intention throughout the following pages is to address the challenges inherent in the analysis of Big Data across the E&P value chain. By definition,

Big Data is an expression coined to represent an aggregation of datasets that are voluminous, complex, disparate, and/or collated at very high frequencies, resulting in substantive analytical difficulties that cannot be addressed by traditional data processing applications and tools. There are obvious limitations working with Big Data in a relational database management system (DBMS), implementing desktop statistics and visualization software. The term *Big Data* is relative, depending on an organization's extant architecture and software capabilities; invariably the definition is a moving target as terabytes evolve into petabytes and inexorably into exabytes. *Business intelligence* (BI) adopts descriptive statistics to tackle data to uncover trends and initiate fundamental measurements; whereas Big Data tend to find recreation in the playgrounds of inductive statistics and concepts from nonlinear system identification. This enables E&P professionals to manage Big Data, identify correlations, surface hidden relationships and dependencies, and apply advanced analytical data-driven workflows to predict behaviors in a complex, heterogeneous, and multivariate system such as a reservoir. Chapter 2 discusses Big Data in more detail and the case studies throughout the book will strive to define methodologies to harness Big Data by way of a suite of analytical workflows. The intent is to highlight the benefits of marrying data-driven models and first principles in E&P.

First Principles

What are *first principles*? The answer depends on your perspective as an inquisitive bystander. In the field of mathematics, first principles reference axioms or postulates, whereas in philosophy, a first principle is a self-evident proposition or assumption that cannot be derived from any other proposition or assumption. A first principle is thus one that cannot be deduced from any other. The classic example is that of Euclid's geometry that demonstrates that the many propositions therein can be deduced from a set of definitions, postulates, and common notions: All three types constitute first principles. These foundations are often coined as *a priori* truths. More appropriate to the core message in this book, first principles underpin the theoretical work that stems directly from established science without making assumptions. Geoscientists have invariably implemented analytical and numerical techniques to derive a solution to a problem, both of which have been compromised through approximation.

We have eased through history starting thousands of years ago when empirical models embraced our thinking to only a few centuries ago when the landscape was populated by theoretical intelligentsia espousing models based on generalizations. Such luminaries as Sir Isaac Newton, Johannes Kepler, and James Clerk Maxwell made enormous contributions to our understanding of Mother Nature's secrets and by extension enabled the geoscientific community to grasp fundamentals that underpin physics and mathematics. These fundamentals reflect the heterogeneous complexity inherent in hydrocarbon

reservoirs. Only a few decades have passed since we strolled through the computational branch of science that witnessed the simulation of complex systems, edging toward the current landscape sculpted by a data-intensive exploratory analysis, building models that are data driven. *Let the data relate the story*. Production data, for example, echo the movement of fluids as they eke their way inexorably through reservoir rocks via interconnected pores to be pushed under natural or subsequently fabricated pressures to the producing wells. There is no argument that these production data are encyclopedia housing knowledge of the reservoirs' characterization, even if their usefulness is directly related to localized areas adjacent to wells. Thus, let us surface the subtle hidden trends and relationships that correlate a well's performance with a suite of rock properties and influential operational parameters in a complex multivariate system. Geomechanical fingerprints washed in first principles have touched the porous rocks of our reservoirs, ushering the hydrocarbons toward their manmade conduits. Let us not divorce first principles, but rather marry the interpretative and deterministic approach underscored by our scientific teachings with a nondeterministic or stochastic methodology enhanced by raw data flourishing into knowledge via data-driven models.

Data-Driven Models

> *The new model is for the data to be captured by instruments or to be generated by simulations before being processed by software and for the resulting information and knowledge to be stored in computers.*[3]
>
> Jim Gray

Turning a plethora of raw upstream data from disparate engineering disciplines into useful information is a ubiquitous challenge for O&G companies as the relationships and answers that identify key opportunities often lie buried in mountains of data collated at various scales in depth as well as in a temporal fashion, both stationary and non-stationary by nature.

O&G reservoir models can be characterized as physical, mathematical, and empirical. Recent developments in computational intelligence, in the area of machine learning in particular, have greatly expanded the capabilities of empirical modeling. The discipline that encompasses these new approaches is called *data-driven modeling* (DDM) and is based on analyzing the data within a system. One of the focal points inherent in DDM is to discover connections between the system state variables (input and output) without explicit knowledge of the physical behavior of the system. This approach pushes the boundaries beyond

conventional empirical modeling to accommodate contributions from superimposed spheres of study:[4]

- *Artificial intelligence* (AI), which is the overreaching contemplation of how human intelligence can be incorporated into computers.
- *Computational intelligence* (CI), which embraces the family of neural networks, fuzzy systems, and evolutionary computing in addition to other fields within AI and machine learning.
- *Soft computing* (SC), which is close to CI, but with special emphasis on fuzzy rules-based systems posited from data.
- *Machine learning* (ML), which originated as a subcomponent of AI, concentrates on the theoretical foundations used by CI and SC.
- *Data mining* (DM) and *knowledge discovery in databases* (KDD) are aimed often at very large databases. DM is seen as a part of a wider KDD. Methods used are mainly from statistics and ML. Unfortunately, the O&G industry is moving toward adoption of DM at a speed appreciated by Alfred Wegener as the tsunami of disparate, real-time data flood the upstream E&P value chain.

Data-driven modeling is therefore focused on CI and ML methods that can be implemented to construct models for supplementing or replacing models based on first principles. A machine-learning algorithm such as a neural network is used to determine the relationship between a system's inputs and outputs employing a training dataset that is quintessentially reflective of the complete behavior inherent in the system.

Let us introduce some of the techniques implemented in a data-driven approach.

Soft Computing Techniques

We shall enumerate some of the most prevalent and important algorithms implemented across the E&P value chain from a data-driven modeling perspective. Three of the most commonplace techniques are artificial neural networks, fuzzy rules-based systems, and genetic algorithms. All these approaches are referenced in subsequent chapters as we illustrate applicability through case studies across global O&G assets.

Artificial Neural Networks

ANNs show great potential for generating accurate analysis and predictions from historical E&P datasets. Neural networks should be used in cases where mathematical modeling is not a practical option. This may be due to the fact that all the parameters involved in a particular process are not known and/or the

interrelationship of the parameters is too complicated for mathematical modeling of the system. In such cases a neural network can be constructed to observe the system's behavior striving to replicate its functionality and behavior.

ANNs (Figure 1.1) are an adaptive, parallel information processing system that can develop associations, transformations, or mappings between objects or data. They are an efficient and popular technique for solving regression and classification issues in the upstream O&G industry. The basic elements of a neural network are the neurons and their connection strengths or weights. In a supervised learning scenario a set of known input–output data patterns are implemented to train the network. The learning algorithm takes an initial model with some prior connection weights (random numbers) and applies an updating algorithm to produce final weights via an iterative process. ANNs can be used to build a representative model of well performance in a particular reservoir under study. The data are used as input–output pairs to train the neural network. Well information, reservoir quality data, and stimulation-related data are examples of input to an ANN with production rates describing the various output bins. Since the first principles required to model such a complex process using the conventional mathematical techniques are tenuous at best, neural networks can provide explicit insight into the complexities witnessed between formation interactions with a stimulation process such as a hydraulic fracture strategy or an acidizing plan. Once a reasonably accurate and representative model of the stimulation processes has been completed for the formation under study, more analysis can be performed. These analyses may include the use of the model in order to answer many *what-if* questions that may arise. Furthermore, the model can be used to identify the best and worst completion and stimulation practices in the field.

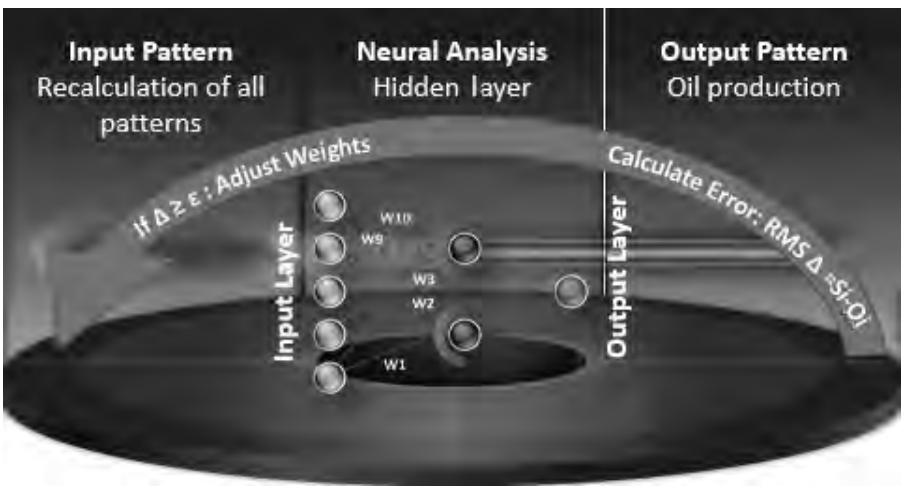


Figure 1.1 Artificial Neural Network

Genetic Algorithms

Darwin's theory of survival of the fittest,[5] coupled with the *selectionism* of Weismann[6] and the genetics of Mendel, have formed the universally accepted set of arguments known as the evolution theory.

Evolutionary computing represents mechanisms of evolution as key elements in algorithmic design and implementation. One of the main types of evolutionary computing is the *genetic algorithm* (GA) that is an efficient global optimization method for solving ill-behaved, nonlinear, discontinuous, and multi-criteria problems.

It is possible to resolve a multitude of problems across the spectrum of life by adopting a searching algorithm or methodology. We live in a world overcome by an almost unlimited set of permutations. We need to find the best time to schedule meetings, the best mix of chemicals, the best way to design a hydraulic fracture treatment strategy, or the best stocks to pick. The most common way we solve simple problems is the *trial-and-error* method. The size of the search space grows exponentially as the number of associated parameters (variables) increases. This makes finding the best combination of parameters too costly and sometimes impossible. Historically, engineers would address such issues by making smart and intuitive estimates as to the values of the parameters.

We could apply an ANN to provide output bins (e.g., 3, 6, 9, and 12 months cumulative production) based on the input to the network, namely, stimulation design, well information, and reservoir quality for each particular well. Obviously, only stimulation design parameters are under engineering control. Well information and reservoir quality are part of Mother Nature's domain. It is essential to implement adjuvant data quality workflows and a suite of *exploratory data analysis* (EDA) techniques to surface hidden patterns and trends. We then implement the genetic algorithm as a potential arbitrator to assess all the possible combinations of those stimulation parameters to identify the most optimum combination. Such a combinatory set of stimulation parameters is devised as being for any particular well (based on the well information and reservoir quality) that provides the highest output (3, 6, 9, and 12 months' cumulative production). The difference between these cumulative values from the optimum stimulation treatment and the actual cumulative values produced by the well is interpreted as the production potential that may be recovered by (re)stimulation of that well.

Fuzzy Rules-Based Systems

How does the word *fuzzy* resonate with you? Most people assign a negative connotation to its meaning. The term *fuzzy logic* in Western culture seems both to realign thought as an obtuse and confused process as well as to imply a

mental state of early morning mist. On the other hand, Eastern culture promotes the concept of coexistence of contradictions as it appears in the Yin-Yang symbol, as observed by Mohaghegh.[7]

Human thought, logic, and decision-making processes are not doused in Boolean purity. We tend to use vague and imprecise words to explain our thoughts or communicate with one another. There is an apparent conflict between the imprecise and vague process of human reasoning, thinking, and decision making and the crisp, scientific reasoning of Boolean computer logic. This conflict has escalated computer usage to assist engineers in the decision-making process, which has inexorably led to the inadequacy experienced by traditional artificial intelligence or conventional rules-based systems, also known as *expert systems*.

Uncertainty as represented by fuzzy set theory is invariably due to either the random nature of events or to the imprecision and ambiguity of information we analyze to solve the problem. The outcome of an event in a random process is strictly the result of chance. Probability theory is the ideal tool to adopt when the uncertainty is a product of the randomness of events. Statistical or random uncertainty can be ascertained by acute observations and measurements. For example, once a coin is tossed, no more random or statistical uncertainty remains.

When dealing with complex systems such as hydrocarbon reservoirs we find that most uncertainties are the result of a lack of information. The kind of uncertainty that is the outcome of the complexity of the system arises from our ineptitude to perform satisfactory measurements, from imprecision, from a lack of expertise, or from fuzziness inherent in natural language. Fuzzy set theory is a plausible and effective means to model the type of uncertainty associated with imprecision.

Exploratory wells located invariably by a set of deterministic seismic interpretations are drilled into reservoirs under uncertainty that is invariably poorly quantified, the geologic models yawning to be optimized by a mindset that is educated in a data-driven methodology.

Fuzzy logic was first introduced by Zadeh,[8] and unlike the conventional binary or Boolean logic, which is based on crisp sets of "true" and "false," fuzzy logic allows the object to belong to both "true" and "false" sets with varying degrees of membership, ranging from 0 to 1. In reservoir geology, natural language has been playing a very crucial role for some time, and has thus provided a modeling methodology for complex and ill-defined systems. To continue the stimulation optimization workflow broached under "artificial neural networks," we could incorporate a fuzzy decision support system. This fuzzy expert system uses the information provided by the neural networks and genetic algorithms. The expert system then augments those findings with information that can be gathered from the expert engineers who have worked on that particular field for many years in order to select the best (re)stimulation candidates. Keep in

mind that the information provided to the fuzzy expert system may be different from formation to formation and from company to company. This part of the methodology provides the means to capture and maintain and use some valuable expertise that will remain in the company even if engineers are transferred to other sections of the company where their expertise is no longer readily available. The fuzzy expert system is capable of incorporating natural language to process information. This capability provides maximum efficiency in using the imprecise information in less certain situations. A typical rule in the fuzzy expert system that will help engineers in ranking the (re)stimulation candidates can be expressed as follows:

IF the well shows a high potential for an increase 3-, 6-, 9-, and/or 12-month cumulative production

AND has a plausible but moderate pressure

AND has a low acidizing volume

THEN this well is a good candidate for (re)stimulation.

A *truth-value* is associated with every rule in the fuzzy expert system developed for this methodology. The process of making decisions from fuzzy subsets using the parameters and relative functional truth-values as rules provides the means of using approximate reasoning. This process is known to be one of the most robust methods in developing high-end expert systems in many industries. Thus it is feasible to incorporate fuzzy linguistic rules, risk analysis, and decision support in an imprecise and uncertain environment.

EVOLUTION FROM PLATO TO ARISTOTLE

Aristotle's sharp logic underpins contemporary science. The Aristotelian school of thought makes observations based on a bivalent perspective, such as black and white, yes and no, and 0 and 1. The nineteenth century mathematician George Cantor instituted the development of the set theory based on Aristotle's bivalent logic and thus rendered this logic amenable to modern science.[9] Probability theory subsequently effected the bivalent logic plausible and workable. The German's theory defines *sets* as a collection of definite and distinguishable objects.

The physical sciences throughout medieval Europe were profoundly shaped by Aristotle's views, extending their influence into the Renaissance, to be eventually revised by Newtonian physics. Like his teacher Plato, Aristotle's philosophy aims at the universal. Aristotle, however, finds the universal in particular things, which he calls the *essence of things*, while Plato finds that the universal exists apart from particular things, and is related to them as their prototype or exemplar. For Aristotle, therefore, philosophic method implies the ascent

from the study of particular phenomena to the knowledge of essences, while for Plato philosophic method means the descent from knowledge of universal forms (or ideas) to a contemplation of particular imitations of these. In a certain sense, Aristotle's method is both inductive and deductive, while Plato's is essentially deductive from *a priori* principles.

If you study carefully the center of Raphael's fresco entitled *The School of Athens* in the Apostolic Palace in the Vatican, you will note Plato, to the left, and Aristotle are the two undisputed subjects of attention. Popular interpretation suggests that their gestures along different dimensions are indicative of their respective philosophies. Plato points vertically, echoing his Theory of Forms, while Aristotle extends his arm along the horizontal plane, representing his belief in knowledge through empirical observation and experience.

Science is overly burdened by Aristotle's laws of logic that is deeply rooted in the fecund Grecian landscape diligently cultivated by scientists and philosophers of the ancient world. His laws are firmly planted on the fundamental ground of "X or not-X"; something *is* or it *is not*. Conventional Boolean logic influences our thought processes as we classify things or make judgments about things, thus losing the fine details or plethora of possibilities that range between the empirical extremes of 0 and 1 or true and false.

DESCRIPTIVE AND PREDICTIVE MODELS

There are two distinct branches of data mining, *predictive* and *descriptive/exploratory* (Figure 1.2), that can turn raw data into actionable knowledge. Sometimes you hear these two categories called *directed* (predictive) and *undirected* (descriptive). Predictive models use known results to develop (or train or estimate) a model that can be used to predict values for different data. Descriptive models describe patterns in existing data that may be found in new data. With descriptive models, there is no target variable for which you are striving to predict the value. Most of the big payoff has been in predictive modeling when the models are operationalized in a real-world setting.

Descriptive modeling involves clustering or segmentation that is essentially the lumping together of similar things such as wells, rock mechanics, or hydraulic fracture strategies. An association is a relationship between two measured quantities that exhibits statistical dependency.

Descriptive modeling techniques cover two major areas:

1. Clustering
2. Associations and sequences

The objective of clustering or segmenting your data is to place objects into groups or clusters suggested by the data such that objects in a given cluster tend to be similar to each other in some sense and objects in different clusters

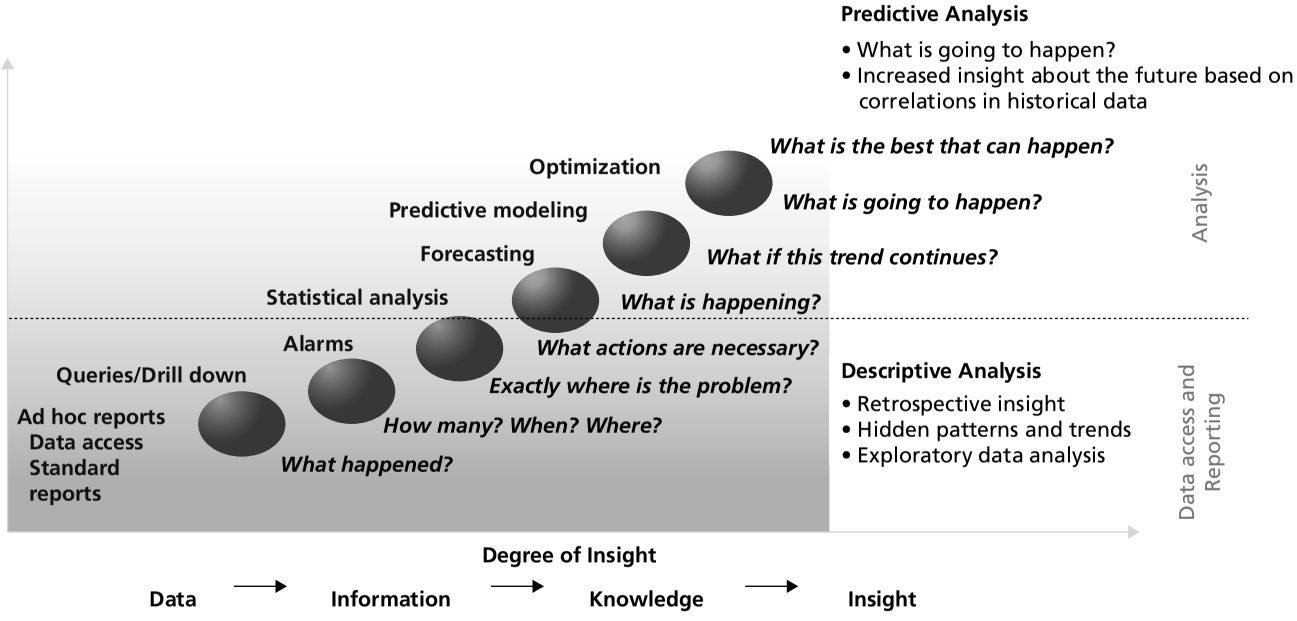


Figure 1.2 Analytics Lifecycle Turning Raw Data into Knowledge

tend to be dissimilar. The term *association* intimates an expansive relationship as opposed to the more limited *correlation* that refers to a linear relationship between two quantities. Thus in quantifying the values of parameters in O&G the term *association* is invariably adopted to underline the non-causality in an apparent relationship.

Predictive modeling appears in two guises:

1. Classification models that predict class membership
2. Regression models that predict a number

There are four main predictive modeling techniques detailed in this book as important upstream O&G data-driven analytic methodologies:

1. Decision trees
2. Regression
   a. Linear regression
   b. Logistic regression
3. Neural networks
   a. Artificial neural networks
   b. Self-organizing maps (SOMs)
4. K-means clustering

Decision trees are prevalent owing to their inherent ease of interpretation. Also they handle missing values very well, providing a succinct and effective interpretation of data riddled with missing values.

An advantage of the decision tree algorithm over other modeling techniques, such as the neural network approach, is that it produces a model that may represent interpretable English rules or logic statements. For example:

> If monthly oil production-to-water production ratio is less than 28 percent and oil production rate is exponential in decline and OPEX is greater than $100,000, then stimulate the well.

With regression analysis we are interested in predicting a number, called the *response* or *Y* variable. When you are doing multiple linear regressions, you are still predicting one number (*Y*), but you have multiple independent or predictor variables trying to explain the change in *Y*.

In logistic regression our response variable is categorical, meaning it can assume only a limited number of values. So if we are talking about binary logistic regression, our response variable has only two values, such as 0 or 1, on or off.

In the case of multiple logistic regressions our response variable can have many levels, such as low, medium, and high or 1, 2, and 3.

Artificial neural networks were originally developed by researchers who were trying to mimic the neurophysiology of the human brain. By combining

many simple computing elements (neurons or units) into a highly interconnected system, these researchers hoped to produce complex phenomena such as intelligence. Neural networks are very sophisticated modeling techniques capable of modeling extremely complex functions.

The main reasons they are popular are because they are both very powerful and easy to use. The power comes in their ability to handle nonlinear relationships in data, which is increasingly more common as we collect more and more data and try to use that data for predictive modeling.

Neural networks are being implemented to address a wide scope of O&G upstream problems where engineers strive to resolve issues of prediction, classification or control.

Common applications of neural networks across the E&P value chain include mapping seismic attributes to reservoir properties, computing surface seismic statics, and determining an optimized hydraulic fracture treatment strategy in exploiting the unconventional reservoirs.

THE SEMMA PROCESS

SEMMA[10] defines *data mining* as the process of **S**ampling, **E**xploring, **M**odifying, **M**odeling, and **A**ssessing inordinate amounts of data to surface hidden patterns and relationships in a multivariate system. The data-mining process is applicable across a variety of industries and provides methodologies for such diverse business problems in the O&G vertical as maximizing well location, optimizing production, ascertaining maximum recovery factor, identifying an optimum hydraulic fracture strategy in unconventional reservoirs, field segmentation, risk analysis, pump failure prediction, and well portfolio analysis.

Let us detail the SEMMA data-mining process:

- **S**ample the data by extracting and preparing a sample of data for model building using one or more data tables. Sampling includes operations that define or subset rows of data. The samples should be large enough to efficiently contain the significant information. It is optimum to include the complete and comprehensive dataset for the Explore step owing to hidden patterns and trends only discovered when all the data are analyzed. Software constraints may preclude such an ideal.
- **E**xplore the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and insightful ideas that insinuate hypotheses worth modeling.
- **M**odify the data by creating, selecting, and transforming the variables to focus the model selection process on the most valuable attributes. This focuses the model selection process on those variables displaying significant attributes vis-à-vis the objective function or target variable(s).

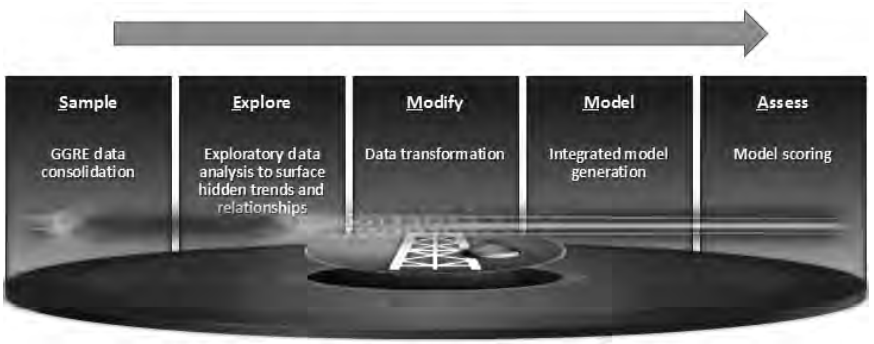


Figure 1.3 SEMMA Process for Data-Mining Workflows

- **M**odel the data by using the analytical techniques to search for a combination of the data that reliably predicts a desired outcome.
- **A**ssess the data by evaluating the usefulness and reliability of the findings from the data-mining process. Compare different models and statistically differentiate and grade those models to ascertain optimum range of probabilistic results, delivered under uncertainty.

It is important to remember that SEMMA (Figure 1.3) is a process, not a methodology. As such, SEMMA is fully compatible with various data-mining methodologies in the IT industry.

HIGH-PERFORMANCE ANALYTICS

High-performance analytics enable O&G companies to be more nimble and confident in their decision-making cycles as they engage in new ventures, generating new value from a tsunami of data. The most challenging fields can be quickly assessed, generating high-impact insights to transform their operations.

With high-performance analytics you can achieve the following:

- Attain timely insights requisite to making decisions in a diminishing window of opportunity.
- Surface insights that once took weeks or months in just hours or days to accelerate innovation.
- Uncover precise answers for complex problems.
- Identify unrecognized opportunities for growth.
- Achieve much improved performance.

In the age of Big Data, O&G companies depend on increasingly sophisticated analysis of the exponential growth in volumes and varieties of data collected at even more frequent rates across the siloed geoscientific community. The velocities of data, coming from intelligent wells equipped with downhole

sensors, are adding enormous pressures on upstream thinking. How can we extract maximum knowledge and cultivate optimized information from raw data? How can we impose quality control workflows that filter noise and outliers, impute missing values, and normalize and transform data values? We must strive to yield a robust collection of disparate data in readiness for both the deterministic and stochastic workflows. It is important to understand that the teachings underpinning the philosophy of this book are not deflecting the traditional interpretations so ingrained by our geophysics, geology, petroleum, and reservoir engineering institutions, but simply emphasizing an important supplement based on the data yielding their hidden secrets. A hybrid approach is optimum, marrying both schools of thought.

In-Memory Analytics

In-memory analytics enable analytical workflows on Big Data to solve complex upstream E&P problems in an unfettered manner. You can also explore solutions to problems you have never even considered due to computing environment constraints.

In-memory analytics scale to your business needs, providing concurrent, in-memory, and multiuse access to data, no matter how big or small. The software is optimized for distributed, multi-threaded architectures and scalable processing, so requests to run new scenarios or complex analytical computations are handled blazingly fast.

It behooves O&G upstream geoscientists to implement in-memory analytics technologies to perform analyses that range from data exploration, visualization, and descriptive statistics to model building with advanced algorithms.

When it comes to the most common descriptive statistics calculations, SQL-based solutions have a number of limitations, including column limits, storage constraints, and limited data-type support. In addition, the iterative nature of EDA and data-mining operations, such as variable selection, dimension reduction, visualization, complex analytic data transformations, and model training, require multiple concurrent passes through the data: operations for which SQL and relational technology are not well-suited.[11]

As an example of the power behind an in-memory analytical architecture, look at the simple heat map in Figure 1.4. Invariably you would send data back to the front-end reporting tools to serially perform complex calculations. But when huge amounts of computations are needed to analyze and produce information, bottlenecks can occur. Implementing in-memory technology performs the calculations on the server, on the fly and in parallel. As a result, computations are very fast because you are not moving large amounts of data elsewhere for processing. Processing can take place on the

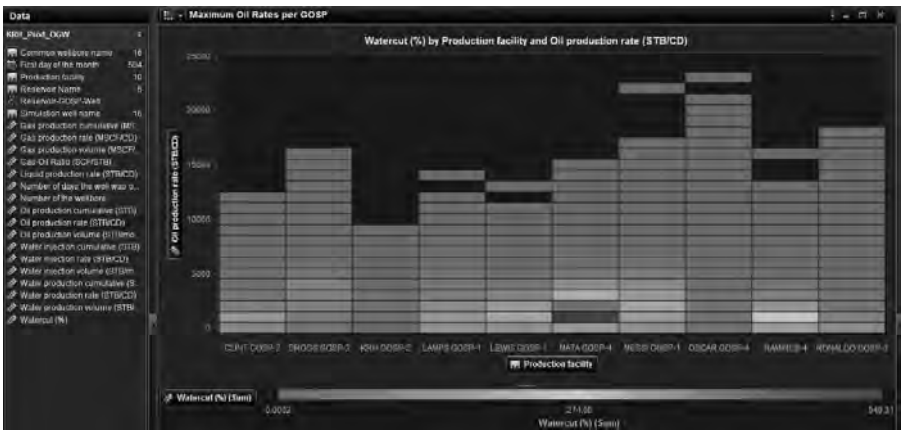


Figure 1.4 Heat Map Highlighting Gas–Oil Separation Plants (GOSPs) and Associated Water Cut

analytic server with the thin results sent back to the client for presentation, rather than for computation.

In-Database Analytics

In-database analytics can execute within database engines using native database code. Traditional processing may include copying data to a secondary location, and the data are then processed using E&P upstream products. Benefits of in-database processing include reduced data movement, faster run-times, and the ability to leverage existing data warehousing investments.[12]

In-database analytics invariably cover two key areas:

1. Develop new products that provide access to and process extant functions within the database.
2. Enhance existing products to leverage database functionality.

In-database processing is a flexible, efficient way to leverage increasing amounts of data by integrating select upstream technology into databases or data warehouses. It utilizes the *massively parallel processing* (MPP) architecture of the database or data warehouse for scalability and better performance. Moving relevant data management, analytics, and reporting tasks to where the data reside is beneficial in terms of speed, reducing unnecessary data movement and promoting better data governance. For upstream decision makers, this means faster access to analytical results and more agile and accurate decisions.

Oil companies operate in a competitive and changing global economy, and every problem has an opportunity attached. Most organizations struggle to manage and glean insights from data and utilize analytic results to improve performance. They often find analytic model development, deployment, and

management to be a time-consuming, labor-intensive process, especially when combined with excessive data movement and redundancy.

In-database processing is ideal for two key scenarios. The first scenario is for *Big Data enterprise analytics*, where the sheer volume of the data involved makes it impractical to repetitively copy them over the network. The second scenario is in complex, organizationally diverse environments, where varying business communities need to share common data sources, driving the need for a centralized enterprise data warehouse. Oil companies should implement corporate data governance policies to promote one single version of the truth, minimizing data inconsistency and data redundancy, and aligning data access needs to common business usage.

Grid Computing

As data integration, analytics, and reporting capabilities grow in strategic importance and encompass increasing numbers of users and larger quantities of data, the ability to cost-effectively scale a business analytics system to gain operational flexibility, improve performance, and meet peak demands using grid computing becomes a competitive advantage.

Grid computing enables O&G companies to create a managed, shared environment to process large volumes of data and analytic programs more efficiently. It provides critical capabilities that are necessary for today's business analytics environments, including workload balancing, job prioritization, high availability and built-in failover, parallel processing and resource assignment, and monitoring.

A grid manager provides a central point for administering policies, programs, queues, and job prioritization to achieve business goals across multiple types of users and applications under a given set of constraints. IT can gain flexibility and meet service levels by easily reassigning computing resources to meet peak workloads or changing business demands.

The presence of multiple servers in a grid environment enables jobs to run on the best available resource, and if a server fails, its jobs can be seamlessly transitioned to another server; providing a highly available business analytics environment. High availability also enables the IT staff to perform maintenance on specific servers without interrupting analytics jobs, as well as introduce additional computing resources without disruption to the business.

Grid computing provides critical capabilities that are necessary for O&G business analytics environments, including:

- Workload management and job prioritization
- High availability
- Parallelization of business analytics jobs for improved performance

Workload management allows users to share resources in order to most effectively balance workload and meet service levels across the enterprise. Business analytics jobs benefit by having workflows execute on the most appropriate resource and multiuser workload is balanced within the grid to enable the optimum usage of resources. Grid computing provides the capability to prioritize jobs, which enables critical jobs to start immediately instead of waiting in a queue. Low-priority jobs can be temporarily suspended to enable critical jobs to be immediately processed.

Grid computing provides standardized workload management to optimally process multiple applications and workloads to maximize overall throughput. In addition, grid computing can parse large analytics jobs into smaller tasks that can be run, in parallel, on smaller, more cost-effective servers with equal or better performance than seen on large symmetric multiprocessor (SMP) systems. Parallelization of upstream analytics jobs enables O&G companies to improve processing speeds by orders of magnitude and deliver exceptional improvements in analyst productivity.

Reservoir simulation programs are best suited for parallel processing owing to potentially large datasets and long run-times.

By combining the power of workload management, job prioritization, and high availability, grid computing accelerates performance and provides enterprises with more control and utilization of their business analytics environment.

THREE TENETS OF UPSTREAM DATA

The three tenets of upstream data are:

1. Data management
2. Quantification of uncertainty
3. Risk assessment

These are key issues in petroleum exploration and development. Oil companies are being forced to explore in more geologically complex and remote areas to exploit deeper or unconventional hydrocarbon deposits. As the problems become too complex in areas of intrinsically poor data quality, and the cost associated with poor predictions (dry holes) increases, the need for proper integration of disciplines, data fusion, risk reduction, and uncertainty management becomes very important. Soft computing methods offer an excellent opportunity to address issues, such as integrating information from various sources with varying degrees of uncertainty, establishing relationships between measurements and reservoir properties, and assigning risk factors or error bars to predictions.

Data Management

We discuss in Chapter 2 the methodologies that underpin data management in the upstream. It is paramount to emphasize the corporate benefits behind automated and semi-automated workflows that enable seamless data aggregation, integration of disparate datasets from siloed engineering disciplines, and the generation of analytical data warehouses (ADWs) in preparation for advanced analytical processes.

With the advent of Big Data in the upstream we are witnessing an explosion of data from downhole sensors in intelligent wells distributed across DOFFs. It is becoming even more essential to implement a concrete enterprise data management framework to address some of the current business issues spawned by an O&G company's critical asset: data.

- Data disparity across systems
- Organizational silos with different data
- Multiple customer views
- The need to access unstructured data within your systems
- Overwhelming growth in data volumes

Quantification of Uncertainty

Do you think the quantification of uncertainty across the E&P value chain has improved over the last few years? And has this progress translated into condensed and more effective decision-making cycles? The answer to the first question is a demonstrative "Yes," but the answer to the second is a qualified "No."

What is happening? Uncertainty quantification is not an end unto itself; removing or even reducing uncertainty is not the goal. Rather the objective is to make a good decision, which in many cases requires the assessment of the relevant uncertainties. The O&G industry seems to have lost sight of this goal in its good-faith effort to provide decision makers with a richer understanding of the possible outcomes flowing from major decisions. The industry implicitly believes that making good decisions merely requires more information. To counter this, let us explore a decision-focused uncertainty quantification framework that will aid in the innovation of better decision-making tools and methodologies. We shall discuss quantification of uncertainty as a common theme threaded through several case studies describing advanced analytics and soft computing techniques.

Risk Assessment

Risk assessment enables decisions under uncertainty to post a risk analysis either through a risk ranking of hazard reduction strategies or through comparison to

target risk levels and cost-benefit analysis. *Risk* can be defined as the product of the consequences of the potential hazard times the probability of occurrence of scenarios. After the risk is calculated, the results must be compared to either governmental or company criteria to determine if the risk is tolerable. This means that the risk is at a level people are generally willing to accept.

EXPLORATION AND PRODUCTION VALUE PROPOSITIONS

If you imagine the analytical algorithms or techniques as the atoms in a molecular structure (Figure 1.5) held together by covalent analytical methodologies or workflows, you get a sense of how the analogy emphasizes the seamless connectivity of soft computing and nondeterministic approaches to aggregate the functions across the E&P value chain that are invariably performed in geoscientific silos.

Oil companies are striving to attain the hidden knowledge in their key asset: data. These data are exploding in volume, velocity, and variety as real-time data from intelligent wells across DOFFs supplement historical interpretations and generated datasets. It is paramount to gain insight from these multiple datasets and enable engineers and stakeholders to make faster more accurate decisions under uncertainty. By marrying the traditional deterministic and interpretive workflows with a data-driven probabilistic set of analyses, it is possible to predict events that result in poor reservoir or well performance or facility failures. Building predictive models based on cleansed historical data and analyzing in real-time data streams, it is now feasible to optimize production. Controlling costs and ensuring efficient processes that impact positively HSE and resource usage are key benefits that fall out of analytical methodologies.

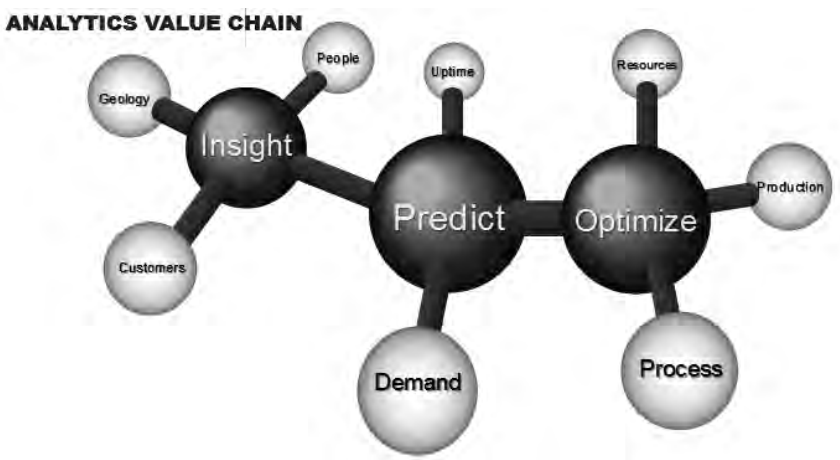


Figure 1.5 E&P Value Propositions

When we think about the lifecycle of an asset, a field or well, there is a business decision that must take place for each phase. That decision must have commercial value and that intrinsic value can be attained by enriching the interpretation from 3D immersive visualization workflows with data-driven models.

Exploration

You could be entering a new play and doing exploration to generate prospects, striving to gain insight from seismic data and to locate exploratory wells in increasingly complex reservoirs.

Appraisal

The appraisal phase of petroleum operations immediately follows successful exploratory drilling. You need to appraise the commercial quantities of hydrocarbons and mitigate risks while drilling delineation wells to determine type, shape, and size of field and strategies for optimum development.

Development

The development phase of petroleum operations occurs after exploration has proven successful, and before full-scale production. The newly discovered oil or gas field is assessed during an appraisal phase, a plan to fully and efficiently exploit it is created, and additional wells are usually drilled. During the development stage a drilling program with optimized completion strategies is enacted as additional wells are located for the production stage. Surface facilities are designed for efficient O&G exploitation. Do we have to consider water production? What cumulative liquid productions do we anticipate? These are some of the questions we need to answer as we design those surface facilities.

Production

The production phase occurs after successful exploration and development during which hydrocarbons are exploited from an oil or gas field. The production phase necessitates efficient exploitation of the hydrocarbons. We have to consider HSE and maintenance schedules. Is the production of hydrocarbons maximized for each well? How reliable are short- and long-term forecasts?

Enhancement

Lastly, enhancement maintains optimal production based on a business decision as to whether an asset is economically viable. How do we identify wells that are ideal candidates for artificial lift? When and how do we stimulate a candidate well?

A well enhancement is any operation carried out on an oil or gas well, during or at the end of its productive life, that alters the state of the well and/or well geometry, provides well diagnostics, or manages the production of the well. There are several techniques traditionally implemented to enhance well production that are categorically termed *enhanced oil recovery* (EOR) or *improved oil recovery* (IOR) artificial lift processes.

OILFIELD ANALYTICS

The oilfield analytics (OA) framework (Figure 1.6) proposes a simple and flexible structure to position data-driven methodologies across the E&P value chain. The profiles of the main actors/stakeholders are then easily associated with the following tenets.

Oilfield Data Management

A robust and consistent suite of data is of utmost importance for successful and credible advanced analytical methodologies. A stable and flexible data management platform is a prerequisite to any soft computing analysis. Once the foundation is established, multiple analytical data marts can be spawned from the master data management (MDM) environment. Over 70 to 80 percent of time is consumed by managing and organizing the upstream data, and with the evolving explosion of data, both historical and real-time, it is a growing business issue in the O&G industry to ensure data integrity. Data flows to/from service providers' popular interpretation tools and products must be integrated into any chosen architecture for a comprehensive solution. We describe O&G data management in Chapter 2.

Oilfield Exploration Analysis

Seismic data are now becoming pivotal as 3D and 4D surveys are accelerated across green- and brownfields. In addition to the customary seismic wavelet data processing, it is becoming more important to fully appreciate the seismic attributes of which there are hundreds and build a seismic data mart for advance analysis. Soft computing methodologies that map seismic attributes to reservoir properties are incredibly important as a means to define more credible and reliable reservoir characterization definitions that underpin field (re) development, supplementing and integrating well logs. The spatial integrity of large areal reservoirs requires high-resolution seismic and a more in-depth understanding of those seismic attributes that can identify both stratigraphic and structural traps. It is essential to attain accuracy, fidelity, and integrity for a seismic data cube containing pre- and/or post-stack traces having been processed with traditional workflows and algorithms such as true amplitude recovery, deconvolution, migration, filtering, and scaling as well as static and

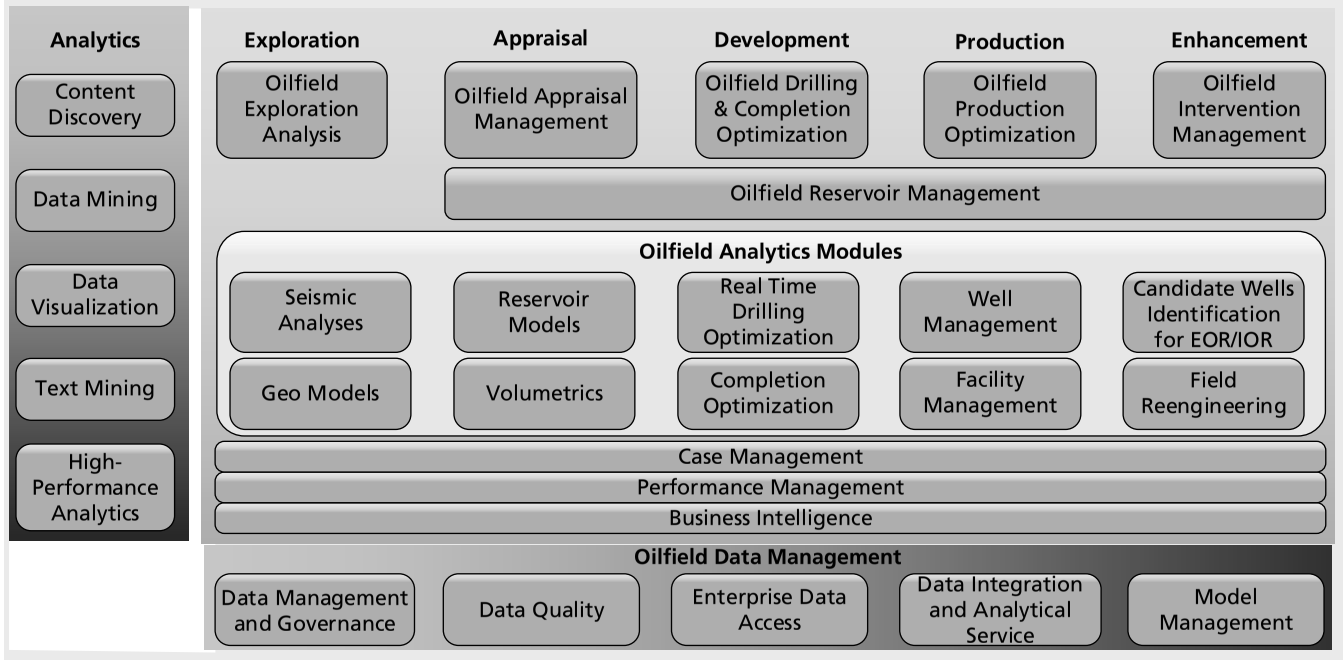


Figure 1.6 Potential Oilfield Analytics Framework

velocity analysis. This is so because these traces are the precursors to the single trace analysis and attribute-derived methodologies described in Chapter 3.

Exploratory drilling is the next step, using drilling rigs suitable for the respective environment (i.e., land, shallow water, or deep water). We cover drilling and completion optimization in Chapter 5.

Oilfield Appraisal Management

The process to characterize the reservoir(s) of a potential or mature field encapsulates the analysis of large datasets collated from well tests, production history, and core analysis results, enhanced by high-resolution mapping of seismic attributes to reservoir properties. It is imperative to capture the more subtle observations inherent in these datasets, to comprehend the structure of the data. Invariably, geostatistical methods can be implemented to accurately quantify heterogeneity, integrate scalable data, and capture the scope of uncertainty. However, between 70 and 80 percent of allotted time for any reservoir characterization study worth its investment should be concentrated on EDA. As an overture to spatial analysis, simulation, and uncertainty quantification, EDA ensures consistent data integration, data aggregation, and data management, underpinned by univariate, bivariate, and multivariate analysis. It is important to visualize and perform descriptive and inferential statistics on upstream data.[13]

If hydrocarbons have been found in sufficient quantities, the development process begins with the drilling of appraisal wells in order to better assess the size and commerciality of the discovery.

It is of paramount importance to undertake the study of uncertainties from engineering design parameters and flow characteristics of the rocks that are not accessible from seismic data. We explore reservoir characterization in Chapter 4.

Oilfield Drilling and Completion Optimization

The target of many operators, particularly in the unconventional assets, is to determine the variables that impact the key performance metric of cost per foot drilled. Other focus areas could be on spud to total depth (TD) and costs of drilling and casing horizontal wells. By using historical drilling data, it is feasible to quantitatively identify best and worst practices that impact the target. The intent is that these insights will improve future drilling operations in unconventional plays and potentially in conventional fields. Advanced analytical methodologies would ultimately develop a predictive model that would provide early warning of deviations from best practices or other events that will adversely impact drilling time or costs. Chapter 5 details some data-driven case studies and workflows to optimize drilling and completion strategies.

The vision of real-time drilling optimization is achievable. Advanced analytical techniques can be applied to gauge how real-time data are analyzed relative to past performance/events to predict downhole tool failures and the ability to do immediate root-cause activities and implement real-time solutions. Some of the benefits gained from establishing consistent drilling methodologies that compare real-time drilling data against previous trends include:

- Avoidance of potential nonproductive time (NPT), by predicting a failure, such as a positive displacement motor (PDM) failure owing to excessive vibration, or in high-pressure and -temperature environments where tool/equipment longevity can be predicted.
- Geo-steering: Able to make real-time adjustments to the well-bore trajectory to achieve maximum reservoir contact based on real-time updates to the geo-modeling data.
- Able to make real-time drilling parameter changes (i.e., weight on bit [WOB], torque on bit [TOB], and flow rate).
- Prevent blowouts: multivariate iterative process to analyze pressures such as formation, mud, and drilling fluid pressures.

Oilfield Reservoir Management

The *reservoir management* component of OA is the nursery or inception of the *digital oilfield of the future* (DOFF). The E&P arena in all O&G companies is taking nontraditional approaches to the more traditional DOFF activities of production optimization and real-time drilling. Ultimately it is a grand design to realize a DOFF and promote the development or evolution of an *analytical center of excellence* (ACE) or event solution center that is at the core of global activities in the upstream environment. Chapter 6 discusses soft computing techniques in reservoir management as well as detailing the requisite steps to establish an ACE.

At the crux of reservoir management are the technologies and methodologies underpinned by advanced analytics that are requisite for a multi-skilled and multidisciplinary control center that enables every aspect of development and production to be handled remotely. When presented with a continuous stream of reservoir, well facilities, and pipeline information, geoscientists and engineers must have automated systems to analyze the data, and help them formulate effective responses to changing surface and subsurface conditions, and the means to implement these responses in real time.

Digital oilfield workflows automate the processes for collection, verification, and validation of the right data so the right people have it at the right time in the right context. However, as investments and capabilities continue to grow, particularly with respect to smarter wells and smarter assets, O&G

companies have new opportunities to turn this information into actionable insights from these efforts. While traditional fit-for-purpose analytic tools function very well for the purpose they were originally designed for, these tools struggle to manage the sheer data growth. And smarter assets produce challenges in how to manage the tremendous growth in data volumes, both structured and unstructured. However, new technologies in real-time analytic processing, complex event processing (CEP), pattern recognition, and data mining can be applied to deliver value from the asset. Chapter 9's section on the early warning detection system studies event stream processing in an analytical workflow.

Oilfield Intervention Management

Intervention optimization remediates wells that have either formation or mechanical issues. It is important to develop a suite of advanced analytical data-mining workflows that implement soft computing techniques such as principal component analysis (PCA), multivariate analyses, clustering, self-organizing maps (SOM), and decision trees to generate descriptive and predictive models that efficiently identify candidate wells for remediation. Implementing an oilfield performance forecasting module to determine robust and reliable probabilistic forecasts for the complete portfolio of wells in an asset is an essential step. Subsequent to this step, real-time production data can be compared to the type curves determined with 90 percent confidence limits to identify wells suitable for intervention. Chapter 7 covers a suite of probabilistic methodologies for forecasting performance across a well portfolio.

Oilfield Performance Forecasting

Analytical workflows can incorporate a decline curve analysis (DCA) step implementing an oilfield production forecasting workflow to identify short- and long-term forecasts for oil, gas, and water production. Implementing mature forecasting models and first principles such as Arps[14] empirical algorithms, you can estimate accurate well performance and estimated ultimate recovery (EUR) and measure the impact, positive or negative, of well remediation techniques.

Comparing real-time production data rates and type curves against forecasted trends, you can:

- Quickly and efficiently identify those wells that require remediation.
- Segment the field via well profile clustering.
- Ratify from a field, reservoir, or well perspective whether current production falls within confidence intervals of expectation and act accordingly.

Oilfield Production Optimization

Advanced analytical methodologies are applicable to perform multivariate analysis on disparate upstream datasets, both operational and nonoperational, to evaluate and determine those variables that either inhibit or improve well performance. Predictive and descriptive analytical workflows combine to explore the data to surface hidden patterns and identify trends in a complex system. Adopting data-driven models in the following areas enables extensive and efficient insight and significant discovery of influential parameters to address issues that adversely impact production, without relying solely on first principles.

There are many production inhibitors, such as skin damage and sanding, that can be predicted by generating models inferred by EDA. Aggregating and integrating datasets from across the geoscientific silos to produce a robust dataset tailored for specific analytical studies are the foundation for all such studies. Analytical workflows can be implemented to attain the following goals:

- Establish variables that are key production indicators.
- Identify critical parameters and their range of values.
- Automate normalization and remediation of all data for missing and erroneous values.
- Identify objective function (i.e., target variable such as recovery factor, liquid carryover, or cumulative non-zero production over a certain period) and determine sensitivity studies to identify key drivers.

Such workflows can identify key performance drivers, and offer strategies and tactics for well completion methods and optimized hydraulic fracture treatment designs. A probabilistic approach helps quantify uncertainty and assess risk for individual field development plans.

Important results from production performance studies adopting aforementioned workflows embrace an automatic methodology to characterize impairment, classify wells as good or bad candidates for well stimulation, predict performance outcomes of particular operational parameters, and increase production with faster decision-cycles. Chapter 8 details advanced analytical workflows to increase production while Chapters 9 and 10 address the range of models and Big Data workflows respectively.

I AM A . . .

Geophysicist

Geophysicists invariably expend a great deal of their time processing and interpreting seismic data to delineate subsurface structure and to evaluate reservoir quality implementing analytical workflows on pre- and post-stack-derived

datasets. The industry is currently focused on exploiting more challenging hydrocarbon accumulations, necessitating the need to incorporate a more diverse suite of data types such as electromagnetic and microseismic.

Whether you are solving exploration, development, or production challenges, you need software tools and advanced analytical methodologies to enable you to easily evaluate all your structural and stratigraphic uncertainties.

Chapter 3, *Seismic Attribute Analysis*, details some important soft computing case studies with an emphasis on applying stochastic workflows on the evolving array of seismic attributes derived from 3D seismic cubes.

Geologist

Geology has its roots firmly planted in science, but the geologist harbors a latent artistic talent that can orchestrate the wealth of subsurface knowledge to reenact the geological processes of deposition, erosion, and compaction predisposing the hydrocarbon reservoirs. You may be addressing challenges across the full gamut of the E&P value chain, and so you need the flexibility to harness advanced analytical methodologies with 3D immersive visualization to meet the pressing business needs. Chapter 4, *Reservoir Characterization and Simulation*, Chapter 6, *Reservoir Management*, Chapter 7, *Production Forecasting*, and Chapter 8, *Production Optimization*, showcase several case studies that illustrate a suite of nondeterministic workflows that generate data-driven models.

Petrophysicist

Intricate data acquired from intelligent wells represent an important investment. It is essential to capitalize on that investment and garner significant knowledge leading to accurate reservoir characterization. Chapter 4, *Reservoir Characterization and Simulation*, and Chapter 9, *Exploratory and Predictive Data Analysis*, offer case studies and enumerate workflows to enable petrophysicists to determine the volume of hydrocarbons present in a reservoir and the potential flow regimes from reservoir rock to wellbore.

Drilling Engineer

Wells are expensive and complex, especially with the advent of unconventional reservoirs. It is essential to integrate geosciences and drilling knowledge so as to ensure drilling and completion optimization leading to smarter and higher quality wells, improved risk management, and reduced nonproductive time.

Well control is crucial in challenging environments, particularly high pressure–high temperature (HPHT) and deepwater, to preclude operating risks

related to wellbore instability and failure. Chapter 5, *Drilling and Completion Optimization*, discusses analytical workflows and soft computing techniques to provide a thorough understanding of data-driven models to predict real-time drilling issues such as stuck-pipe.

Reservoir Engineer

The O&G industry is inundated with reservoir simulation software to generate a suite of numerical solutions that strive to provide fast and accurate prediction of dynamic behavior. Owing to the array of reservoirs and their inherent complexity in structure, geology, fluids, and development strategies, it is paramount to adopt a top-down workflow that incorporates artificial intelligence and data-driven models. Conventional wisdom assumes that the reservoir characteristics defined in a static model may not be accurate and thus in a history matching workflow can be modified to attain a match. The functional relationships between those characteristics are deemed as constants derived from first principles. However, a reservoir engineer can question the constancy of the functional relationships and adopt an AI and DM methodology that makes no *a priori* assumptions about how reservoir characteristics and production data relate to each other. Chapter 4, *Reservoir Characterization and Simulation*, Chapter 6, *Reservoir Management*, and Chapter 9, *Exploratory and Predictive Data Analysis*, offer some supportive case studies and salient discussion points to enrich a reservoir engineer's toolbox. Chapter 8, *Production Optimization*, discusses methodologies in a case study to optimize hydrocarbon production via maximized well placement strategies implementing data-driven workflows.

Production Engineer

Data-driven models are ideal supplementary methodologies that deliver significant performance improvements across the current E&P landscape that is densely populated by digital oilfields and intelligent wells generating a plethora of disparate raw data. *Data Management* (Chapter 2), integrated production operations, and performance optimization are key goals for any asset. Production engineers can leverage innovative technology capabilities that are scalable and tailored to individual assets. Marry the data-driven and soft computing techniques to traditional models to fully exploit an asset's riches. Chapter 6, *Reservoir Management*, Chapter 7, *Production Forecasting*, and Chapter 8, *Production Optimization*, offer nondeterministic workflows to endorse the soft computing methodology.

Petroleum Engineer

As a petroleum engineer you are concerned with all activities related to oil and gas production. It is imperative to fully capitalize on all available subsurface data to estimate the recoverable volume of hydrocarbons and comprehend a detailed appreciation of the physical behavior of oil, water, and gas within the porous rocks. The low-hanging fruit of the global oilfields have been discovered and gradually depleted. It behooves petroleum engineers to take advantage of the improvements in computer modeling, statistics, and probability analysis as the advent of Big Data across the E&P value chain, and the complexity of subsurface systems, force the industry to adopt a data-driven analysis. Chapter 4, *Reservoir Characterization and Simulation*, Chapter 7, *Production Forecasting*, and Chapter 8, *Production Optimization*, offer data-driven methodologies focused on production optimization. Chapter 9, *Exploratory and Predictive Data Analysis*, discusses soft computing techniques that are integral to developing models based on subsurface data, driving the analytical solution.

Petroleum Economist

It is essential to harness the technical skills of petroleum engineers with the foresight of economists to empower better business decisions in the E&P industry. Chapter 7, *Production Forecasting*, discusses an integrated *decline curve analysis* methodology providing robust short- and long-term forecasts of well performance.

Information Management Technology Specialist

The O&G industry continues evolving to multiple databases, structured (data) and unstructured (documents) data management, web services, and portals, as well as hand-held devices and cloud computing. It is about building on technology advancements to provide an integrated and innovative solution that has real and measurable value to the E&P business. New client problems must now be solved, including the management of disruptive applications and growing data complexity, frequency, and volume. This chapter broaches several topics that impact the IT professional, particularly the sections on high-performance analytics. Chapter 2, *Data Management*, details some of the critical thinking behind the endorsement that data are a key asset, especially in the sphere of data-driven models.

Data Analyst

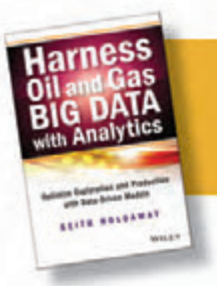
What is a data analyst? The analysis of data is a multifaceted task that incorporates inspection, cleansing, transformation, and modeling of data. In the

world of O&G companies, such a persona adopts artificial intelligence and data-mining techniques on E&P data to pursue knowledge discovery. Essentially this entire book complements the profile of data analysts as they implement exploratory data analysis, descriptive and predictive models through data visualization techniques (Chapter 9) and integrate text analytics (Chapter 10) to garner business intelligence.

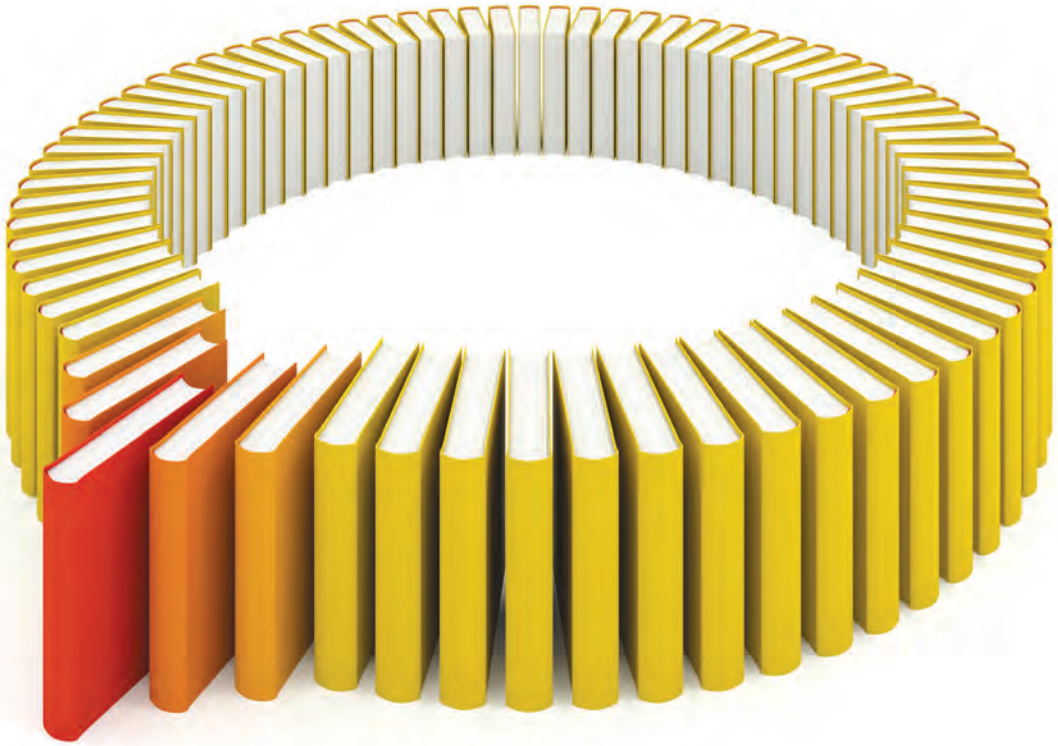
NOTES

1. Bertrand Russell, *The Philosophy of Logical Atomism* (London: Fontana, 1972).
2. J. T. Tukey, *Exploratory Data Analysis* (Reading, MA: Addison-Wesley, 1977).
3. Jim Gray, "A Transformed Scientific Method," based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, January 11, 2007.
4. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Reading, MA: Addison-Wesley, 1989).
5. Charles Darwin, *On the Origin of Species*, 4th ed. (London: John Murray, 1866).
6. A. Weismann, *Essays upon Heredity* (London: Oxford Clarendon Press, 1889).
7. S. Mohaghegh, Virtual Intelligence and Its Applications in Petroleum Engineering, Part 3. *Fuzzy Logic, Journal of Petroleum Technology*, November 2000.
8. L. A. Zadeh, "Fuzzy Sets," *Information and Control* 8, no. 3 (1968): 338–353.
9. G. F. L. P. Cantor, "On a Property of the Collection of All Real Algebraic Numbers," *Journal fur die reine und angewandte Mathematik* 77 (1874): 258–262.
10. SAS Institute defines *data mining* as the process of Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA) large amounts of data to uncover previously unknown patterns.
11. "In-Memory Analytics for Big Data," SAS Institute Inc, White Paper, 2012.
12. Paul Kent, R. Kulkarni, U. Sglavo, "Turning Big Data into Information with High-Performance Analytics," Datanami, June 17, 2013.
13. K. R. Holdaway, "Exploratory Data Analysis in Reservoir Characterization Projects," SPE/EAGE Reservoir Characterization and Simulation Conference, 19–21 October 2009, Abu Dhabi, UAE.
14. J. J. Arps, Analysis of Decline Curves, *Transactions of the American Institute of Mining Engineers* 160 (1945): 228–247.

From *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models* by Keith R. Holdaway. Copyright © 2014, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Harness Oil and Gas Big Data with Analytics*. Full book available for purchase here.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

sas
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0413