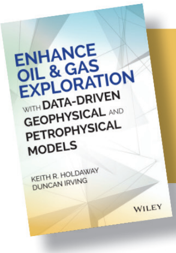
ENHANCE OIL & GAS EXPLORATION

WITH DATA-DRIVEN GEOPHYSICAL AND PETROPHYSICAL MODELS

KEITH R. HOLDAWAY
DUNCAN H. B. IRVING

WILEY

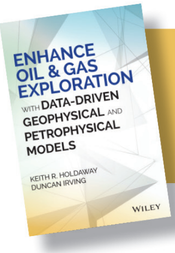


ENHANCE OIL & GAS EXPLORATION WITH DATA-DRIVEN GEOPHYSICAL AND PETROPHYSICAL MODELS

KEITH R. HOLDAWAY
DUNCAN IRVING

WILEY

Contents



CHAPTER 1

Introduction to Data-Driven Concepts

"Habit is habit and not to be flung out of the window by any man, but coaxed downstairs a step at a time."

Mark Twain

INTRODUCTION

Current Approaches

We wish to air some of the more important practical considerations around making data available for data-driven usage. This could be for static, offline studies or for operationalized, online reviews. We introduce the concept of data engineering—how to engineer data for fit-for-purpose use outside the domain applications—and we take the reader from the first baby steps in getting started through to thoughts on highly operationalized data analysis.

A geoscience team will use an extensive collection of methods, tools, and datasets to achieve scientific understanding. The diversity of data spans voluminous pre-stack seismic to single-point measurements of a rock lithology in an outcrop. Modeling approaches are constrained by:

- Size and scarcity of data
- Computational complexity
- Time available to achieve a "good enough" solution
- Cloud computing
- Budget
- Workflow lubrication

It is this last constraint that has proven the largest inhibitor to the emergence of a data-driven approach in exploration and production (E&P). It is a motif for the ease with which data and insight are moved from one piece of software to another.

These constraints have led to a brittle digital infrastructure. This is problematic not only in the individual geoscientific silos but also across the wider domain of E&P. We can potentially exclude a rich array of data types, and restrict innovative methodologies because of the current hardware/software stacks that have evolved symbiotically. The application-centric landscape undermines E&P solutions that strive to integrate multidimensional and multivariate datasets.

It was not meant to be this way. Back when it all began, it was okay for decisions to be made in an expert's head. High-performance computers (HPCs) were power tools that gave the expert better images or more robust simulations, but at the end of the workflow, all that number crunching led to a human decision based on the experience of that human and his or her team of peers. Currently, there is too much riding on this approach.

So, how do we become data-driven if it's hard to get at the data?

Is There a Crisis in Geophysical and Petrophysical Analysis?

There is a movement to adopt data-driven analytical workflows across the industry, particularly in E&P. However, there is an existing group of Luddites providing not constructive criticism but deliberate and subversive rhetoric to undermine the inevitable implementation of data-driven analytics in the industry. It is true data scientists sometimes lack experimental data of a robust nature. How certain are we that we can quantify uncertainties? How can we understand the things that manifest themselves in the real world, in the hydrocarbon reservoirs? They argue that without concrete experimental evidence, theory harbors the risk of retreating into metaphysics. Predictive and prescriptive models are only the source of philosophical discourse. It is tantamount to solving the problem of how many leprechauns live at the end of our garden. Science is not philosophy. Thus, without recourse to experiment, geoscientists play in the realm of pure speculation and march to the metaphysical drumbeat of ancient philosophers. The slide into metaphysics is not always clear. The language of the perplexing mathematical algorithms can mask it. Theoretical physics, especially quantum physics, and the theories that underpin the geosciences and E&P engineering disciplines can be jam-packed with opaque, impermeable, thorny mathematical structures. The Luddites, looking over the soft computing techniques and data-driven workflows, are betrayed into believing that only the high mathematics and classical physical laws must deliver rigor, a wisdom of the absolute, the lucidity of the variance between right and wrong. No doubt there is rigor. But the answers we get depend so much on the questions we ask and the way we ask them. Additionally, the first principles can be applied incorrectly and the business problem unresolved for the engineers asking the questions.

So, there is no crisis unless we wish to create one. The marriage between traditional deterministic interpretation and data-driven deep learning and data mining is a union that when established on the grounds of mutual recognition, addresses an overabundance of business issues.

Applying an Analytical Approach

The premise of this book is to demonstrate the value of taking a data-driven approach. Put simply, if the data could speak for itself, what would you learn beyond what your current applications can tell you?

In the first place, it is the experience of many other industries that statistical context can be established. This could be around testing the validity of an assumed scientific assumption (for example, water flood versus overburden compaction being the cause of a 4D velocity change) or it could be demonstrating whether a set of observations are mainstream or outliers when viewed at the formation, basin, or analog scale.

The current crop of applications:

- Lack the computational platform for scale-out analysis
- Can only consume and analyze data for which they have an input filter

- Are only able to use algorithms that are available in the code base or via their application programming interfaces (APIs)

We discuss in greater detail ahead how to get G&G (geological and geophysical) data into a useable format, but first let us set the vision of what could be plausible, and this takes us into the world of analytics.

What Are Analytics and Data Science?

Analytics is a term that has suffered from overuse. It means many things in many industries and disciplines but is almost universally accepted to mean mathematical and statistical analysis of data for patterns or relationships.

We use this term in customer- and transaction-rich industries, as well as domains where businesses operate on the thinnest of margins. In the UK in the 1950s, the Lyons Tea Company implemented what we now recognize as centralized business intelligence. It was a digital computer that performed analytics across its empire-wide supply chain: thousands of teashops and hundreds of bakeries. Their business analytics grew from their ability to understand and articulate their business processes regarding a data model: a description of the relationships between entities such as customer and inventory items. The team that built this system (called Leo) went on to create similar platforms for other organizations and even sell computing space. This presaged the central mainframes of IBM by a decade, the supply chains of Starbucks by four decades, and the cooperation/competition of computing resources pioneered by Amazon. This history is well documented (Ferry, G., 2010, "A Computer called LEO") and is worth bearing in mind, as we understand how the paradigm applies to the geoscientific domain.

Let us fast-forward to the late 1990s and the evolution of the Internet beyond its academic and military homelands. Data could be collected from across an organization and transmitted into, around, and beyond its conventional boundaries.

This gave businesses no technical reason to avoid emulating Lyons's example of 40 years before, and those that could exploit the ability to process and assimilate their data for business impact pulled ahead of those that proved unwilling or unable to embrace this technical potential. Davenport's "Competing on Analytics" is a mesmerizing overview of this dynamic period in business history (Davenport, Harris, 2007).

As well as the ability to move data around using well-designed and implemented protocols (i.e., via the Internet), the data was generated by:

- Interactions between people and organizations via interfaces such as point-of-sale terminals or ATMs
- Communications between individuals and agencies via web-based services
- The capture of data along a supply chain as goods and materials—or people in the case of travel and hospitality industries—moved around a complex system

Data arising from a transaction could be captured trivially at sufficient quality and richness to enable statistical insight to be gained, often in real time, in the instance of assessing the likelihood that it is someone other than a banking card's owner using it at a given location and time.

Analytics is provisioned by the integration and contextualization of diverse data types. Moreover, it is predicted by timely access to reliable, granular data. If we look to the downstream domains of our industry, this would be real-time access to real-time data about refinery operations and productivity and passing it through to trading desks to enable capacity to be provisioned against spot pricing options.

The economic luxury of $100 oil insulated a lot of the upstream domain from adopting this type of integration. With the growth of factory-style drilling for unconventional plays, development and lifting costs became a major component of the economics. Since 2014, it has become less unusual (but still not mainstream) for drilling engineers to be guided in their quest for best practices. Such guides include analytical dashboards that are the result of combining petrophysical, technical, and operational data in statistical models. Engineers can use such guidance to characterize likelihoods of bit failure or stuck pipe under given geological and operational parameters.

The big surprise from working on such projects is not the willingness of rough-necked senior drillers to embrace such an approach (money, especially saved costs, always talks), but more that the data types in question could be brought together and used in such a manner. This combined an approach that used to be called *data mining* (it's still an appropriate term but is now deeply unfashionable) and soft computing techniques, which currently fall under the definition *data science*.

To a dyed-in-the-wool data miner (and probably a senior drilling engineer), data science is one of those unpleasant necessities of modern life (so it's probably an age-related thing). *Data science* is an umbrella term embracing mathematics, especially statistical expertise, domain understanding, and an intimate knowledge of the domain data and the different format standards. Clearly, this is beyond the capabilities of one single person, hence the widely circulated concept of the data science unicorn.

However, our experiences suggest that such a team should:

- Be configured as small as possible
- Contain a mathematical component that can cope with the physical sciences
- Deal with the worst of formats and the poorest data quality

Data science, done well, has been the difference between liquidity (and at least the next round of venture capital) and history for startups and mega-scale incumbents in many industries in the twenty-first century. It may seem, on the first encounter, to be an ad-hoc, ungoverned approach to working with data and working in general, but it has yielded dividends when applied formally in an organization.

If there is the political will in an organization to accept and act on findings from data science activities, then it will have a quantifiable business impact. Hence, it is reasonable to assume that data science becomes a measurable and valued capability in that organization. It requires a cultural change to provide pervasive impact, but we all must start somewhere, and small bite-sized projects run with a well-constrained scope in an agile manner can yield impactful results. The endpoint is a continuous conveyor of insight generation, through business validation and into operational usage, the DevOps mindset.

As an industry, we are a long way from A-B testing of our processes in the way that online retailers will test different views of their website on statistically sub-selected groups of their clientele to assess session profitability (yes, they do). There is a lot that can be learned about the behavior of many things that we don't think of in population terms (e.g., wells, formations, offset gathers of traces), and the relationships that may exist within and between such logical or statistical groupings.

Meanwhile, Back in the Oil Industry

With this landscape in sight, let us now turn our gaze to our industry. E&P workflows are designed with the goal of delivering high-value insights into the subsurface world. Data is acquired often at high cost and contains information of potentially enormous economic value. While the general types of data have not changed much since the first seismic surveys were performed and the first wells drilled, the scale of acquisition has increased by orders of magnitude.

We are still trying to measure the properties and behaviors of the subsurface and the engineering developments, interventions, and operations that we apply to the subsurface. But, in contrast, the time available to provide insight has shortened from years to months, or even weeks and days. Workflows are compressed in response to fit more agile portfolio decision making and operationalized development and production environments.

However, the business units involved in the upstream domain have hardened into brittle silos with their disciplines, processes, and technological predilections with data compartmentalization. There is an old approach to data curation, with lineage and provenance often missing, and this leads to a fundamental lack of trust in data on the rare occasion that there is the political and physical will to move it from one business silo to another.

With each silo being driven by its key performance indicators (KPIs), they can often be working at odds with each other. The information technology (IT) and operational technology (OT) capabilities in each domain have prevented data, used at operational and tactical levels, from being given enterprise visibility and value. Hence, there is no analytical culture in the upstream domain of our industry. (We often turn to the refining and trading domains as occasional beacons of good practice.)

With no data-driven culture, there is a weak alignment of business challenges across silos and processes, and no analytical capability has emerged at the enterprise scale. The economic upheaval of the 2014/15 price crash stunned the industry and laid bare its inability to respond to challenges at this scale as the underlying processes were so brittle. However, there is a focus emerging on how cost and value can be tied to processes and activities at ever-more granular scales. This is more predominant in the operations and production domains, but the impact is tangible.

The risk is that the same mistakes are repeated. There is a cultural mistrust between the operational business units and the corporate IT teams that should or could support them. This led to the outsourcing of software development and data processing to proprietary systems from data historians to seismic acquisition and processing. This removal of control over algorithms, data platforms and whole architectures in the case of sensor data yielded control over how data can support a business to the service companies and consultancies and is one of the most notable differences between the Oil and Gas industry and the industries mentioned earlier.

We are in peril of echoing the same mistakes by positioning analytics as point solutions that fail to scale or join up with other analytical endeavors. Without a data-driven culture, there is no strategic ownership of an analytics capability, and it is common to see duplication of effort on the IT and the operational sides of the business with competition for human and platform resources. The same service companies are attempting to fill the void by extending their suites of tools with point solutions, which exacerbates the underlying problem that there is no coupled approach to an analytical culture, the tools used for analysis, or the data that provisions the business insight.

We hope that this book shows how seismic, subsurface, and reservoir data can be used to drive business impact. The techniques that we present cover a broad range of geoscientific problems and while they may be useful in and of themselves, it is the approaches and underlying mindset that are the key messages that we wish to convey.

Depressed oil prices have focused primarily on cost controls and on extracting more value from existing workflows.

How Do I Do Analytics and Data Science?

The fundamental difference between data-driven insight generation and conventional techniques is that the former is labor intensive, whereas the latter usually only requires a software subscription. While there is much value to be had from the range and sophistication of applications, they are limited in the range of data that they can assimilate and the scale at which they can perform this adaptation.

Let's conduct an interesting thought experiment to see how cross-functional insight can be achieved using a data science team. There is the obvious need for business sponsorship via the medium of expected business impact. Delivering a new insight of the behavior of property *X* in the context of the property of scenario *Y* is a good place to start. Returning to our driller, it would be insightful to understand the modes of operation to be avoided in each petrophysical and geomechanical context, and base this on experiences gained over several hundred (or thousand) drilling campaigns.

How can this goal be achieved? First, we will consider what resources are needed in a typical endeavor, and when they are deployed along a workflow. At this stage, we concern ourselves with extracting data-driven insights to learn something new. Later we will discuss what we do with this insight and what impact it could have. Will it affect operational processes? Will it drive new technical approaches? Is it something that can be used to generate a rule or parameterize a model? In a continuously evolving deployment, how does the insight become operationalized?

To enable data to be turned into insight, a variety of skillsets is required, which fall into three broad domains:

- The *data domain*—what does the data describe, how is it stored, how can it be accessed, how is it formatted, what does it look like (data demographics and texture—more on this later). We have made this extremely difficult for ourselves in the Oil and Gas industry by keeping everything in application silos and insisting on moving things around using clunky formats specified decades ago (there are good reasons for this, but still).
- The *problem domain*—this is what makes the Oil and Gas industry difficult. The problems are some of the

hardest—up there with moon landings—and require some serious brainpower as well as very sophisticated mathematical algorithms to simulate the processes and dynamics of our space. The algorithms have become so erudite that they now drive stovepipe workflows and it is very tough to put context and insights from other domains into play.

- The *analytical domain* is poorly constrained. We wouldn't be writing a book on data-driven approaches in this industry if it were a well-established discipline. An understanding of statistical methods is a prerequisite. However, the methods that have become conventional in other industries sometimes sit awkwardly and require a practical implementation in the face of ugly data, subtle physical processes, and operational idiosyncrasies that defy characterization.

Each domain contains business-specific challenges as well as technical requirements. Input is needed from the business to ensure effort has validity and impact, as well as providing subject matter expertise.

- The data is sparse in some dimensions and finely sampled in others (e.g., wells, or seismic volumes). It is often costly to acquire but rarely viewed as an asset, or treated as a byproduct of an expensive activity but discarded as digital exhaust. It requires considerable subject matter expertise coupled with data engineering skills ranging from conventional shell scripting through to XML, JSON, and binary parsing.
- The problem space demands a strong understanding of the physical sciences and the mathematics to accompany it if the value is to be generated. The level of mathematics usually implies some level of computing ability in most protagonists, but this is rarely formally learned, or scalable, and it rarely has any statistical component to it.

- Analytics in Oil and Gas has seen a slow adoption rate as analytical and statistical software has found more conducive markets. Many business verticals adopt more consistent data standards, and the business problems appear easier to solve with more tangible and quantifiable results. The key stumbling blocks encountered by the authors in their implementation of analytics have been applying analytics to time series and developing strategies to overcome sparseness in data in one dimension or another. There are also significant computational challenges in driving analytics with upstream data.

What Are the Constituent Parts of an Upstream Data Science Team?

Hence, a team that will work on data-driven projects needs a blend of geosciences, physical sciences, mathematics, statistics, and computer science supported by data architecture. Through experience, it has become clear that a physical sciences background and an average of several years of industry experience is a prerequisite in delivery unless a team is happy to have people learning on the job. This mix of skills and experience is the key ingredient required to perform data science, and it is highly likely that an upstream data scientist would possess most of these skills.

We caution strongly against taking a group of general data science resources and applying them in the upstream domain. There is a definite need to develop strong skills overlap in such a team for productive work to cope with people moving in and out of a project. More specifically, a data science team that does not have significant upstream domain expertise or a mathematical skillset that spans both the physical sciences *and* advanced statistics will fail, generating budgetary and reputational fallout. Correspondingly, there is much to be said for bringing in methods from other industries and problem spaces, but it must be anchored with strong overlap in the data, problem, and analytic dimensions for high impact.

The flipside of this is also problematic. We defined a data-driven study as placing the data and insights of one domain in the context of those from another domain. If a study is too narrow in its scope, it risks being compared unfavorably to well-established applications, algorithms, or workflows. Any new insight will likely be incremental and much groundwork must be applied to the data to establish unique value from the effort derived. Moreover, it will only have business impact at the level where the data is used in isolation. The more data types, the higher the potential impact, as the problem space addresses more of a system or value-chain. Another problem encountered is preconceived notions carried from industry professional training that sometimes lead to acceptance of a process, relationships, or correlations that are not statistically true.

There is a tension or compromise that must be dealt with in repeated data-driven activities. An investment in time and resources is always made in the first encounter with a new data type and problem as a team gets to grips with the data, understands its behavior, and applies appropriate analytical techniques and visualizations. There should be a return on this investment, but at the same time a data science team must retain its objectivity and stay abreast of new approaches and techniques that might be implemented. It is an organizational risk to embed a data science team for an extended period (more than 3–6 months) in any project or domain, and this has long been acknowledged in other industries where "crop rotation" is enforced by strategic KPIs. The upstream space is diverse enough to promote good data science evolution and career paths; individual disciplines such as drilling or reservoir monitoring are too narrow to sustain an intellectually able data science team, however long the wish-list of projects may be.

Strong data science teams are cross-disciplinary, quick to collaborate, business-focused, effective and efficient with the use of technology, and comfortable with failure (it will happen occasionally). It is worth reflecting on whether the upstream domain is currently capable of creating or attracting personnel that could thrive in such teams.

A DATA-DRIVEN STUDY TIMELINE

If you are embarking on a data-driven study for the first time, it is a steep, but exhilarating, learning curve. One of the biggest challenges will be showing value in a specific timeframe. Whether the project is in the commercial or the academic environment, resources are committed as part of an economic consideration and you may only have these resources (people and computational platform) available for a finite period.

The more people are involved in the study, the more dependencies you must deal with on their engagement. Also, there is often the need to ensure that all team members are aligned in progress and vision, so regular review, refocusing, and planning must take place. An agile methodology works well in this context as a good compromise between experimentation and productivity. The three core activities are to:

1. *Learn something new from your data assets:* Is there a pattern, trend or relationship in your data that is telling you something that no one has seen before?
2. *Place that knowledge in a business context:* Does it add value, save cost, change operational process?
3. *Document everything:* The chances are that this will not be the last time you do this. Record code, methods, problems, figures, and reports.

All three areas should be viewed as areas of outcome and should be explicit in the planning, funding, execution, and reporting of a project. At the scale of a large organization, if data-driven methods are being deployed to complement and enhance existing approaches, then a longer-term view on sustainable and scalable implementation must be developed. Smaller studies are the leading edge of this movement and are a way of learning how to perform data-driven techniques, hence the need to understand the business, technical, and process challenges and benefits.

Realistically, a small data-driven study is anything from one to eight weeks. A one-week study is at the same scale as a coding "hackathon." This is an event that has become popular in many organizations where the use of technology is a competitive differentiator. It allows low-risk experimentation with software (and often hardware) by teams of professionals (e.g., mathematical modelers) and business stakeholders (e.g., drilling engineers) to test ideas. This general premise is extended to longer time frames based on considerations around data engineering, projected time to business value, and complexity of analytical tasks. In our experience, six to eight weeks is a typical timeline for a successful discovery project.

For a pure discovery-style study, a simple set of workflow gateways should be constructed, which are typically as follows:

- Pre-project work:
  - Identify use case.
  - Define data.
  - Agree on success criteria.
- Data preparation:
  - Acquire data.
  - Load data to a staging area.
- Work preparation:
  - Understand toolsets.
  - Identify analytical packages.
- Work packages:
  - Execute analytical packages.
  - Continuous documentation.
  - Review, obtain feedback, and plan next package.

- Review:
  - Present to stakeholders.
  - Review and obtain feedback.

The pre-project work could take several weeks of meetings and requests for data until there is agreement across all parties concerned that sufficient data is available to enable outcomes that justify the effort and the resourcing. The data preparation and data engineering are often the most poorly scoped aspects of projects. If a team is meeting a data type for the first time, then it is not unusual for a few hundred person-hours to be devoted to unlocking the structure and behavior of the data; pre-stack seismic would fall into this category. Conversely, parsing a few thousand Log ASCII Standard (LAS) files out into a useable form is often a matter of hours as the format is well understood and tools freely available.

Once the study has been defined and the data loaded and understood, the analytical work can begin. A shared understanding of the potential business questions across the team performing the analysis and any stakeholders is vital, even if it is simply “What is there in my petrophysical data that I haven’t spotted when viewed at the basin scale?” The analytical tools will likely be agreed at this stage, and an understanding of the shape and size of each of the steps should be developed. An agile methodology of planning, sizing, and execution could be applied if the study is to run for more than a few days.

Regular regrouping to ensure alignment of effort is vital, and if running for several weeks, then periodic reviews with stakeholders are necessary to ensure expectations are managed, the value is communicated, and new ideas can be drawn if progress permits. All the time, documentation should be a background activity. There are several tools available that allow code to be stored and shared in online repositories (public or private, e.g., Github). There are simple platforms and a service that teams can use to document their work (e.g., Jupiter, Apache Zeppelin) with working code and statistical algorithms. It is necessary to provide interactive visualizations to help with communicating the outputs from their analysis. This ensures that the insights can live on well beyond the completion of the project as opposed to fossilizing them in PDF and PowerPoint.

What Is Data Engineering?

For this book, we view data engineering as the design and implementation of a data access framework. It covers the extraction of data, metadata, and information from source files and transforming them into a form, view, or analytical dataset to enable data-driven analysis. It should consider governance around security, quality, and lineage and engineering, data reuse, extensibility and scalability in size, complexity, and execution speed may also be considerations. Crucially, it is the cultural bridge between the curational world of subsurface data management and the insight-producing domain of analytics and data science. As in a construction project, where an architect must listen to the client and create a building that at the same time is aligned with a shared vision and meets the customer's needs, so the engineer must execute against these requirements using his or her understanding and experience with the practicalities of the materials to be used.

So, in a data-driven analytics environment, matters could be as simple as lining up a collection of time series samples from disparate data domains along a consistently tested timeline, or they could be challenging as ingesting passive seismic data from hundreds of sensors and extracting patterns and features for operationalized analytics.

If the data-driven study is a one-off activity, then shame on you for lack of vision. Let us explore this road momentarily and then move on. Data must be extracted from a native file format, an application database, or some other form of transfer mechanism (Excel, plain text) and then cleaned, validated, and placed into some structure that allows analysis. This could be a table in a database, a data frame in R or Python, or some custom structure in any one of the many big data number-crunching platforms. The effort has been expended with no value to show.

Before spending the energy, think of the future and consider what might happen if the insights of your data-driven analysis are considered valuable. You will be asked to repeat it with a larger dataset, more sophisticated algorithms, combined with other data types, and more—will you have to repeat all the steps and expend the same effort, or can you repeat, reuse, scale, and extend your efforts with ease? If we answer in the negative, then you need to consider your approach to data engineering.

A Workflow for Getting Started

We offer a set of guiding principles rather than a rigid methodology. The amount of time spent in preparing the data for use and engineering more robust functionality around your work is governed by the quality of the data, the volume and structure of the data, how much integration is required, and how quickly the whole workflow needs to be performed—from one-off to continuously streamed data.

We will stay away from stream processing architectures and remain in the shallow end of the analytics pool in trying to expose data at its most granular level, bring it to a level of quality that is fit for analytics, and apply context using any metadata and by integration with other data. What follows is a roughly linear workflow with the caveat that you should expect iterations through it until you arrive at a dataset and resulting insights that are robust enough to drive a business decision.

Opening the Data

The first practical step is getting hold of the data. Often this is a political challenge as much as a practical one. Confidence must be won, usually with the promise of a stake in the project and a sharing of outcomes. When requesting data, ask for as much as possible. That's not idealism speaking; that's a request for metadata. When you ask for "everything" you can then ignore unimportant data at your leisure; but when asking for the "raw" curves you realize that you forgot to call for the well headers, and so on, you get the idea. You'll understand that you need data about the data (metadata) and some reference data (master data) such as the Master Curve List of well curves or the official stratigraphic terms used by your organization or client. This is your first step on the road to context and hence enlightenment.

In many Oil and Gas companies, it's hard to locate, or sometimes access, the "official" or "correct" version of such data and a gray area exists that is inhabited by people's favorite or most trusted version of a given dataset or reference table. We have all seen examples of this murky world, and checking the veracity of data as it moves from one domain application to the next is one of the major time-sinks in subsurface analysis workflows.

Metadata, Master, and Measurement Data

Metadata is a necessity in the subsurface data world. It is the anchor for all those physical measurements and interpretations. As geoscientists, we like to think that we know exactly when and where all our costly measurements were made. In the real world, it is customary to hear anecdotes about the geodetic baseline and reference ellipsoid being lost when a system is migrated from one database to another as part of technology refresh or when undergoing acquisition by another organization.

Fortunately, our subsurface data managers are smart people even if the systems that they must use are not, and their technical committees created data exchange formats for various data types that have stood the test of several decades (something that nearly all other file formats have failed to do!). Seismic, well, and production datasets all contain data that tell the user (in human-readable text) how to unpack the data and what each field means. At the very least we have a logical framework, if not a spatial or chronological one, from which to set out. If you're lucky, you will then have enough metadata in the form of headers, comments, and accompanying master data that will allow you to place your measurement data in the correct spatial and logical context.

The location of this metadata is well defined (and well adhered to, usually) in formats such as SEG-Y (seismic) and LAS (well log), and what is easily readable by the human eye is now easily assimilated by parsing scripts in a variety of languages. A consistent data science team should find it simple to parse trace headers, well headers, or any other industry metadata to extract necessary information about survey/well names, and common-depth point (CDP)/well header locations. Several open-source projects are now hosted on public code bases such as "Github" to get started.

Data Types

Let us briefly consider the three main classes of data that are encountered in the subsurface domain and reflect on the challenges of each class. Our data is usually a measurement or collection of measurements at a location and a particular time. We take raw physical measurements and perform all manner of cleaning, interpolation, and refactoring to give a better measurement but, raw or synthetic, we are trying to describe the subsurface in space and time. There will also be contextual information contained in the text that may need to be extracted and integrated at scale.

Chronological Data

Beyond single-point measurements, time series are the simplest data types. Typically, they are measurements of the same property at the same location, ideally at regular intervals. If the interval is irregular, then some interpolation and resampling strategy are often required to provide a consistent dataset (computers and more importantly our algorithms prefer regular sampling). Analytics often requires asking a system what is happening at a point in time, or across a discrete window if the concurrence constraint can be justifiably relaxed.

Chronological data is best converted to an International Organization for Standardization (ISO) timestamp data type, which requires some careful parsing, conversion, and concatenation of data that often comes in Julian Day format if a boat has been involved in its acquisition (e.g., seismic data).

As a special case of temporal data, well logs represent a time series, masquerading as a simpler 1D dataset. Remember that well logging is a collection of rock properties sensed at a regular sampling rate while a logging tool is pulled up a borehole. This is then converted to a down-hole depth, but it should be noted that mismatches can occur, and composite logs are not immune to errors creeping in where logging has taken place at different rates by different contractors.

Similarly, seismic imaging data is also a collection of discrete time series windows. The time is two-way travel time, and the data is presented as bunches of time series attached to fixed points on a survey. Pre-stack seismic data is more complex since the time becomes a critical access path when assembling simultaneous events such as a shot gather.

While these last two examples of time series may seem contrived, they illustrate the fact that there needs to be careful thought as to what questions we are asking of our data before we plan how to store and access our data for analysis. The industry standards for data formats were developed for robust and fool-proof *transfer* of data, and not for ad-hoc access to granular data at scale.

Let us consider the simplest case more closely. Imagine that we have a single sensor taking a measurement at frequent and regular intervals (one second, for the sake of this example). Let us assume that we measure a property that changes rapidly enough that we need to sample it every second, and is part of an operational control system that we can tap into for logical reasons. We would like to understand the longer term (e.g., week–month) scale behavior of our property. We will address the types of analysis that are appropriate for this later; suffice to say we will have a very long, thin dataset. We shall quickly fill a spreadsheet beyond the point that a human brain and eye can extract meaningful insight, and moreover, it will present an indexing challenge if stored as a single physical file.

For accurate time series, it may be necessary to resample to a standard timestamp for analysis or—better—use a time series database. This is an emerging class of databases that allows ranges of historical data to be extracted and compared even when events do not fall on exact timestamps. This was until recently not engineered into mainstream databases, but the rise of the Internet of Things (IOT) agenda and its industrial equivalent have seen considerable investment in time-based analytical capabilities.

For 1D data that happens to be a discrete time series—seismic traces and well logs—there are still decisions to be made about how to access data in the time series. However, this needs to be balanced with how each measurement is indexed. Adding additional indexes for easting, northing, acquisition time (or survey identifer for 4D seismic), parameter name (for well logs), and offset (for pre-stack seismic) require extra storage. Such storage needs have to be justified as regards the value of provisioning so many ways of accessing the data for instant analysis.

Spatial Data

Spatial data presents its class of problems, which are dealt with efficiently in many other publications. For a thorough grounding in the theory, we recommend *Spatial Data Modelling for 3D GIS* (Abdul-Rahman and Pilouk, 2008) as a starting point. Most analytical approaches should support some spatial representation of relationships between data. It is possible to decompose any 2D or 3D dataset into its most granular form where the analytical value requires it. As with time series data, it then becomes a series of design decisions on how to provision access at scale and a degree of performance.

The ability to access geospatial subsets of contiguous data is often used for specific geological data. We shall see in the petrophysical use cases that it is more often the case that there are relationships at play in the data that are hidden from us. We insist on storing and manipulating the data as a 2D or 3D unit rather than letting the data show us the dimensions in which it contains the most information.

As long as data governance is strong, that is, you don't lose the information (aka master data) about coordinate reference systems, reference ellipsoids, and the reference datum of a dataset, then it is possible to take more abstract relationships in the data and project them faithfully back into our physical world.

Textual Data

Textual data in this sense refers to documents, as well as comment fields within applications, that contain written information that can be incorporated into a data-driven study. This is more typically to add context to numerical data rather than as a source of statistical or measurement data in its own right. Text analysis is a massive area of research and we introduce it here to signpost its applicability and low barrier to entry.

At its simplest text analysis looks for words and clusters of words in a document. The end goal is to distill a document into a reduced vocabulary that can be extended to other data types. Examples are equipment inspection and operational notes, geological interpretation, or observations during seismic acquisition. Commonly occurring words are identified and then tuples of two, three, and four words are inspected for deeper context (e.g., sandstone, fine-grained sandstone, fine sandstone).

Adjustments for spelling can be made, and eventually, a reduced vocabulary can be derived. Where the data quality is high, it has been possible to develop predictive models of varying validity, and more sophisticated statistical approaches in this area are discussed by Chen et al. (2010).

Hence, the transformation is not a simple geometric or structural manipulation of the data. It is the extraction of information contained in the data—information that is then becoming a property or attribute to provide context about a location, area, event, period, and so on. Hold this thought as we progress through feature engineering.

Making Your Data Useable

A critical capability in your data science skillset is the ability to understand when you have a data quality issue. It is straightforward to inspect data and see where non-numerical characters occur where you expect to see a number. However, it requires increasing degrees of sophistication to establish what a range of allowable values are; what the expected precision should be; or whether a blank, a null, an NaN (not a number), or a value (e.g., −999 in LAS files) should be respected and resolved by imputation. Data should be engineered such that rules are developed and applied consistently—in agreement with a domain expert where necessary—to ensure that insight is robust from the first pass and that any future work builds on strong foundations.

Dealing with poor-quality data by removing values can lead to another problem—data sparsity. Is there enough information contained in a dataset for meaningful insight? Sparsity also requires rules. If data is missing, then should the last value be used, or a null value, or an interpolated value? If it is interpolated, then what approach should be utilized?

If data is prone to error—mainly instrumental error—then statistical methods should be used to smooth it. Such approaches could be as simple as a running mean, applied by passing a window along a dataset, to more sophisticated statistical techniques, including *statistical process control* and *moving-window* *principal component analysis* (PCA). This is entering the territory of time-series analysis, and there are a multitude of techniques that can be deployed. As in the real world, it is feasible to clean data too much; a smoothing filter or overly aggressive interpolation will remove the very detail and variability that contains the information required for the analysis. This is where an iteration through domain expertise is vital to ensure that data and insights are statistically valid while retaining as much information in the source data as possible. Good data engineering will then allow this to be built up and appended efficiently.

The Road to Data Science Perfection

We apologize for the tongue-in-cheek heading, as perfection is something that we see as a long way off at the time of writing—and something we hope to see changing in our geoscience world quite soon. We hope that the following reasoning—based on many analytical projects executed by the authors—shows the value of getting the data preparation workflow as robust as possible before embarking on what may seem to a business stakeholder as the high-value activity of analysis. Our experience gives rise to a cautionary and measured approach.

Data Profiling

Data profiling goes beyond basic error checking to tell us something about the behavior or character of the data. Simple metrics such as its variability or standard deviation are useful, as are ranges, means, and medians. This is elementary statistics, but it is a mathematical domain that many geoscientists may not have encountered for some time. We illustrate this with well-logging data in Chapter 3, concerning petrophysical data, where statistical profiling at the formation level is a simple and efficient metric that can be stored alongside the raw data.

At a more generalized level, consider an infinitely long time series of data. There is a signal contained in the data, and for a thought experiment it can be a simple harmonic signal with a lot of background noise. Imagine you are listening to a flute being played through a thin wall and the window is open so traffic noise is there in the background and must be filtered out. Suppose a flute plays a concert pitch note of A above middle C (440 Hz) and the sound is being sampled 1 ms; then you acquire 1000 bytes per second. We are assuming that you can describe the amplitude of your microphone in a 32-bit byte (this is reasonable, so don't think too hard about it).

Now, what if the pitch of the flute rises and falls for whatever reason. As a subject matter expert (i.e., you're the neighbor who listens to this all day and every day) you observe that this rising and falling off a constant note is not abrupt but changes slightly over a period of many seconds. Continuing in this imaginary world and pretending that we are set the task of monitoring the pitch of the flute over the course of several hours, let's say that we only have a spreadsheet for the purpose. It becomes apparent that we will overcome the size limitations within a matter of minutes if we attempt to record sound intensity as measured by a microphone every millisecond.

This is the idea of profiling surfaces. It is not the raw data, but the frequency (or pitch) of the data at any given instant that interests us. Moreover, as the frequency is varying slowly, it only requires samples every ten seconds, per our subject matter expertise. Hence we now have one number—pitch—that describes the data at a given period, say every second. This provides all relevant information but reduced bin volume by three orders of magnitude. Signal processing may be required to extract this from the background noise, but judicious filtering will achieve this and leave the necessary signal intact.

Now consider the background noise. There is the steady hum of traffic and potentially the odd aircraft. We notice that as an airplane passes overhead, the music becomes livelier with sequences of notes being played. Our flutist neighbor is an aircraft fan and the sight of a plane lifts the music. If we are calculating the standard deviation (a measure of the variability) in the data through time, we see that it has a higher range and standard deviation of its pitch at such times and a linear regression performed by the data scientist backs this up readily.

Is there a way of characterizing what our neighbor plays when a plane is sighted? There is, and this is feature engineering. We see that there is a characteristic pattern of the notes, for instance, and we always hear a progression of the same four notes when sighting a plane. This progression is a pattern or feature that we should keep as we now have an early warning that an aircraft is approaching and we should close our windows to avoid the deafening noise. We have just performed a feature engineering thought experiment.

Feature Engineering

Using the example of a set of notes to define a feature, it should be evident that there are many applications to this in geosciences data. This could be the sea state or tidal behavior in maritime operations, a litho-facies in a well log, acoustic facies in a seismic survey, or a dynamic reservoir effect seen in production response.

Put simply, feature engineering allows a geoscientist to identify and capture all the exciting aspects of a dataset that would have been sketched and noted in a field notebook in the physical world. Data science techniques lead to the prominent features, and it is the subject matter expert—in this book, the geoscientist—who then assigns context or meaning or otherwise. Even better is to allow other data to provide the context.

It is beyond the scope of this chapter to review specific use cases or algorithms as data, and mathematical approaches display so much diversity across the E&P domain. The purpose of this extended thought experiment is to show that it is at least equally useful—if not more so—to present the raw data in an accessible and well-curated form. We must also keep profile data and key features of the data alongside it. It is the features that provide the analytical hooks: "Where do I hear this sequence of notes?" becomes "Where do I see these facies?"

Analytical Building Blocks

These features and statistical parameters become analytical building blocks. As familiarity with your data evolves, so the statistical methods become more sophisticated and more abstract. It is at the feature engineering level that the subject matter expertise needs to bake in a lot of the scientific relationships in the data. At risk of laboring the point, the data quality and data preparation strategies are vital if the features are to be useful for longstanding analytical deployment in a business context.

In many cases, it is these features rather than the raw data itself that become the starting point for regression and machine learning (ML) algorithms. If the subject matter expert has validated that the features and statistical metrics contain sufficient information about the data, then there is a much higher likelihood of running successful ML workflows on the data at scale as opposed to developing an architecture for data mining, statistical processing, and ML workflows of granular data at scale.

It should become the norm for dimension reduction and characterization steps to be performed, and the benefits of good data governance become apparent if these features are to be reused across several studies at the scale of a large organization like an oil company or the academic community. Equally, well-crafted and -engineered features may well become the intellectual property of an organization if the competitive benefit can be derived from their ownership. This question is likely to vex the industry for several years hence!

IS IT INDUCTION OR DEDUCTION?

> *"Though this be madness, yet there is method in it."*
>
> Hamlet

The objective behind *deep learning* (DL) is without question the art form that is *induction*. How does it differentiate from *deduction*?

- *Induction:* the cerebral path from factual minutiae to general principles.
- *Deduction:* traveling in the opposite direction to induction, it follows the reasoning from the general to the specifics or from cause to effect.

Through induction in deep learning we are striving to make sense of the big data accumulated across disparate engineering sources: data from multiple sensors that record what has happened in the system under investigation. We then draw grandiose conclusions as we identify trends and patterns in these datasets. Essentially, we are reverse-engineering Mother Nature's physical laws and first principles. As Polonius observed in Hamlet: it is the uncovering of the *method* in the *madness*.

We must make assumptions during our induction process since there are many irrational behaviors noted in the data. There is no such thing as a perfect understanding of the way a system works. Hence the learning method is based on simple assumptions that are an intelligent way to identify patterns that are useful in our DL methodology.

It seems the Oil and Gas industry is riddled with engineers and geoscientists who are tightly anchored to the deductive reasoning espoused by Aristotle and posited by Hobbes in his arguments with Wallis in the 1600s. If the analytical methodology is deficient in rigor and invariably takes you down the road to paradoxes, then it is *not* precise and scientifically acceptable. Zeno's famous example of contradictions, illustrating the celebrated "Achilles and the Tortoise" enigma, adds substance to contemporary attitudes against inductive logic. Why was the

Royal Society initially distrustful of mathematics when established in England during the 1600s? Because of the illustrious founder, fellows such as William Ball, Sir Robert Moray, and subsequently Wallis and Wren revered experimental science. Of course, much of the argument for inductive as opposed to deductive reasoning was born from the liberal ideals of the day that fought the Jesuit iron fist that seemed to be choking society in the seventeenth century.

However, Wallis, who stood as the single mathematician in the hallowed corridors of the Royal Society, took up the mantle to merge mathematics with the life force of the Society. He claimed, "Mathematical entities exist not in the imagination but reality." In short, he supported the experimental methodology that has since evolved into today's data-driven analytical workflows under the banner of *data science*. This is in stark contrast to the Euclidean perspective of geometry. Wallis argued that constructing geometrical objects from the first principles is contrary to the natural world order where such geometry exists in Mother Nature. He stated that the study of geometrical figures was analogous to examining the geologic strata in the subsurface. Like Wallis, the modern geoscientists should rely more on inductive logic and hence integrate data-driven methodologies within the rigorous context of the first principles. Why? Because simple deduction stifles new ideas and induction paves the way to revolutionary ideas being seeded as we toy with new perspectives that reflect the reality of nature. Without induction, Newton would not have invented calculus as a branch of mathematics to address the issues generated by the "method of indivisibles," a suspicious technique per the Jesuits who poured scorn on those striving for an explanation of "infinitesimals." So, let us not reject induction when applying a reasoned and logical approach to solving reservoir characterization or simulation across the geophysical and petrophysical sciences. Stuck in deduction just to adhere to the first principles will deflect from the realistic, even if probabilistic, results garnered from a data-driven methodology born in induction.

The Oil and Gas E&P activities are at an intersection. There is an increasing strain between the accepted and prevailing image of mathematics as an assemblage of eternal and unchanging truths and its actual implementation in the global reservoirs replete with uncertainties, frustrations, and failures. Do we as geoscientists wish to perpetuate, like the Jesuits in the 1600s, the appearance of academic infallibility at the expense of exploring new ground and innovative techniques? Remember theoretical and practical advancements in all sciences are invariably engendered from bizarre ideas.

With that in mind, let us uncover the probabilistic insights from some advanced data-driven techniques applied across the geophysical and petrophysical sciences when applied to data generated in these silos.

REFERENCES

Abdul-Rahman, Alias, and Morakot Pilouk, *Spatial Data Modelling for 3D GIS* (2008). DOI: 10.1007/978-3-540-74167-1.

Amir, Alexander, "Infinitesimal: How a Dangerous Mathematical Theory, Shaped the Modern World," *Scientific American*/Farrar, Straus & Giroux (April 2014).

Boman, Karen, "Study: Low Oil Price Gives Industry Chance to Pursue Digital Transformation," *Rigzone*, May 12, 2015, www.rigzone.com/news/oil_gas/a/138503/Study_Low_Oil_Price_Gives_Industry_Chance_to_Pursue_Digital_Transformation, accessed July 27, 2015.

Chen J., Z. Li, and B. Bian, "Application of Data Mining in Multi-Geological-Factor Analysis." In: Cai Z., C. Hu, Z. Kang, and Y. Liu (eds.), "Advances in Computation and Intelligence," ISICA 2010, *Lecture Notes in Computer Science*, vol. 6382, Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-16493-4_41.

Davenport, T. H., Harris, J. G., *Competing on Analytics: The New Science of Winning*, 2007.

Ferry, G., *A Computer called LEO: Lyons Tea Shops and the World's First Office Computer*, 2010.

Jacobs, Trent, "High-Pressure/High-Temperature BOP Equipment Becoming a Reality," *Journal of Petroleum Technology*, 67, no. 7, www.spe.org/jpt/article/6707-ep-notes-5/, accessed July 27, 2015.

Kane, Gerald C. et al., "Strategy, Not Technology, Drives Digital Transformation," Deloitte University Press, Summer 2015, http://52.7.214.27/articles/digital-transformation-strategy-digitally-mature/, accessed July 27, 2015.

Slaughter, A., G. Bean, and A. Mittal, "Connected Barrels: Transforming Oil and Gas Strategies with the Internet of Things" (2015), https://dupress.deloitte.com/content/dam/dup-us-en/articles/iot-in-oil-and-gas-industry/DUP-1169_IoT_OilGas.pdf.

Teradata, "Reduce Operational Complexity to Cut NPT," www.teradata.com/industry-expertise/oil-and-gas/, accessed July 27, 2015.

About the Authors

Keith R. Holdaway is advisory industry consultant and principal solutions architect at SAS, where he helps drive implementation of innovative oil and gas solutions and products. He also develops business opportunities for the SAS global oil and gas business unit that align SAS advanced analytics from exploratory data analysis and predictive models to subsurface reservoir characterization and drilling/production optimization in conventional and unconventional fields.

Prior to joining SAS, Holdaway was a senior geophysicist with Shell Oil, where he conducted seismic processing and interpretation and determined seismic attributes in 3D cubes for soft-computing statistical data mining.

* * *

Dr. Duncan H. B. Irving has been a leading consultant in oil and gas for Teradata since 2010. Prior to that he researched and instructed in petroleum geoscience at the University of Manchester, and provided freelance upstream data management consulting. Throughout his career he has worked on data acquisition, integration, and analytics around oil reservoir, subsurface, and sensor data in data centers, in extreme field conditions, and in general upstream workflow and data management.

Duncan has led and supported projects across the oil and gas and wider manufacturing industries, consulting at strategic and operational levels. Projects in these industries span scientific, technical, and business domains, and Duncan stays wide in his approaches, for example, marrying modern data science paradigms to longstanding supercomputing-driven workflows.

He has slowly swapped Perl for Python and PowerPoint for data art. He has a PhD in glacial geophysics, publishes and speaks regularly on oil industry data and analytics challenges, and enjoys being at the forefront of the emerging analytical ecosystem in upstream oil and gas.

Ready to take your SAS® and JMP® skills up a notch?

Be among the first to know about new books, special events, and exclusive discounts.
support.sas.com/newbooks

Share your expertise. Write a book with SAS.
support.sas.com/publish

sas.com/books
for additional books and resources.

sas®
THE POWER TO KNOW.®