When the Data Grows Faster Than the Day

AI turns massive archives into usable insight.

By MEHER GAJULA, KEYLA GONZALEZ, BEN LASSCOCK, SATHIYA NAMASIVAYAM, ALTAY SANSAL and ALEJANDRO VENCIANO of TGS

Oil and gas exploration grapples with an awkward equation: the data grows, but the day does not. Cloud platforms have changed availability, with petabytes of SEGY, LAS files, and reports stored in systems that virtually never go down. AI now changes the rate and scale of what we can learn from this data. With columnar, cloud-native formats like open source, MDIO and Parquet, our scientific data is not only stored, but also computable and teachable.

The new inflection is the combination of generative AI and domain-specific foundation models, working together with a new generation of automated AI agents. These systems don't merely search; they connect, linking survey seismic data with other metadata, processing histories, entitlements, and interpretations. Creating business content becomes shorter, clearer, and, crucially, repeatable at an enterprise scale.

The Bottlenecks We All Know

Most organizations recognize the pattern. Seismic, well logs, acquisition and processing reports, interpretations and business or contract documents live in separate systems, often with distinct owners and access policies. As an example, entitlement risk is woven through the fabric: what we can use, where and when depend on legal nuances scattered across PDFs, slide decks and inboxes.

Object storage was intended to simplify access, but legacy formats and folder structures don't align with modern retrieval patterns.

Even when everything is "in the cloud," it can still be effectively offline for practical analysis. Meanwhile, a great deal of domain judgment on how to read a vintage text header or how a survey's lineage affects interpretability resides with experts who carry that knowledge in their heads.

Assembling a fit-for-purpose dataset to answer even a single question can take weeks or months. The backlog grows; the workforce is stretched. We have more data than ever, yet the decisions we need to make continue to arrive faster.

Why AI Finally Matters

The shift of the past two years has been measurable. Generative AI agents now read and reason over more than a hundred thousand documents in hours, retrieve what is relevant to a specific question and summarize the implications for action. They translate natural language into SQL, vector search and API calls, orchestrating retrieval without asking specialists to write one-off scripts.

Text large language models can classify reports and interpret figures; image-capable systems can describe the content of geophysical plots. But the breakthrough for our domain lies in foundation models trained on the raw scientific data itself: pre- and post-stack seismic volumes, well logs and associated metadata. Pretraining builds deep, transferable representations that accelerate interpretation and reduce the setup cost for common tasks. The same

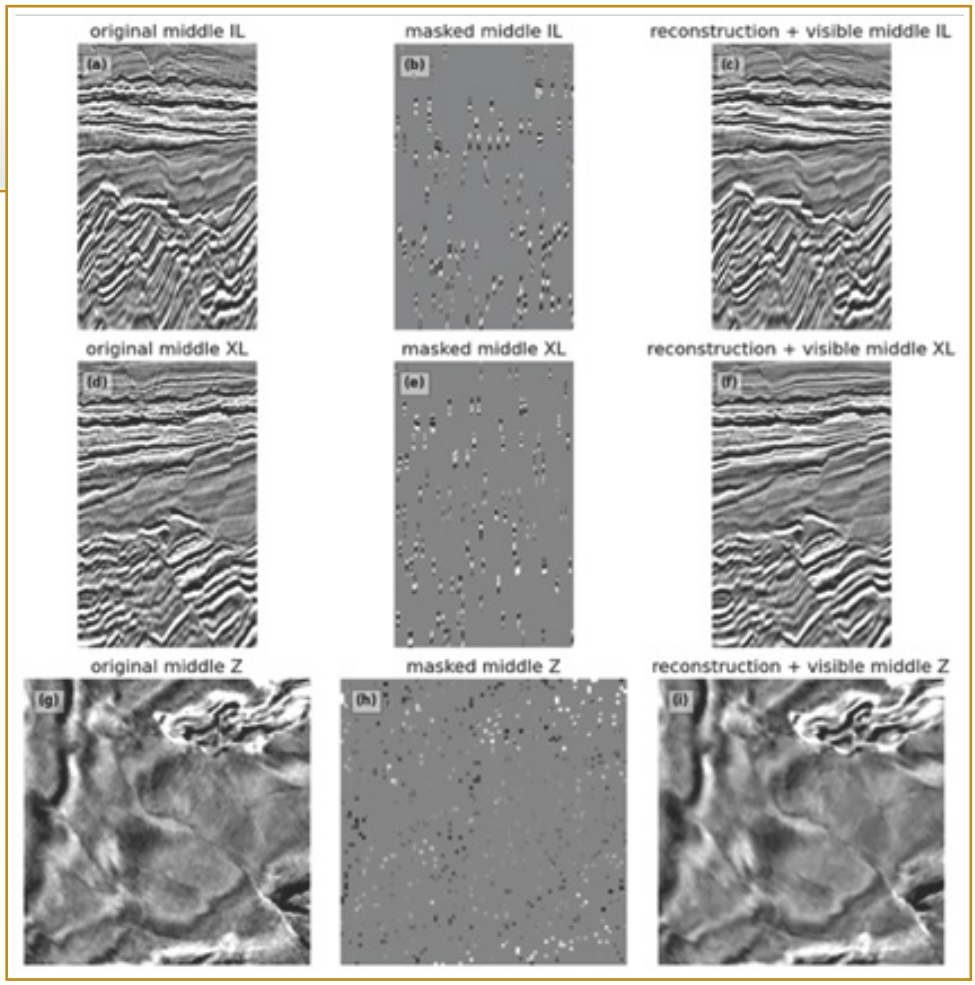

Figure 1. (Left) 3-D seismic input, shown as inline, crossline and depth slices. (Middle) Randomized sampling of the inputs data, with 90 percent of data hidden in training. (Right) A reconstructed image from the 10-percent available data.

pattern is visible in adjacent fields, from Earth observation to biosensing: pretrain broadly, adapt specifically, deliver faster insight. Google's DeepMind and AlphaEarth Foundations, for example, help map our planet in unprecedented detail. Underpinning all of this are cloud-native, analysis-ready formats, such as MDIO and Parquet. They are the "glue" that enables generative AI, vector retrieval, and foundation models to operate on large, shared datasets with predictable performance and governance.

Case Studies

▶ Pre-training a seismic foundation model: Pre-training scientific data builds a model that is forced to learn representations of the specific data to solve a task. In this example (figure 1), the model is trained to recover withheld portions of a 3-D seismic volume from the remaining 10 percent of observed data. Its performance is evaluated by how well it reconstructs the original volume, not by generating new data. Similar pre-training strategies can be applied to well data or other scientific domains, with related examples seen in wearable sensor studies [3].

The result is a model that has learned representations that generalize across basins and surveys, drastically reducing the setup time for downstream tasks such as salt mapping or fault detection. ▶ Formation tops interpretation: A well foundation model, trained on 1.1 million North American well logs, exposes a promptable task for top prediction. By mining a database of historical

To next page ▶

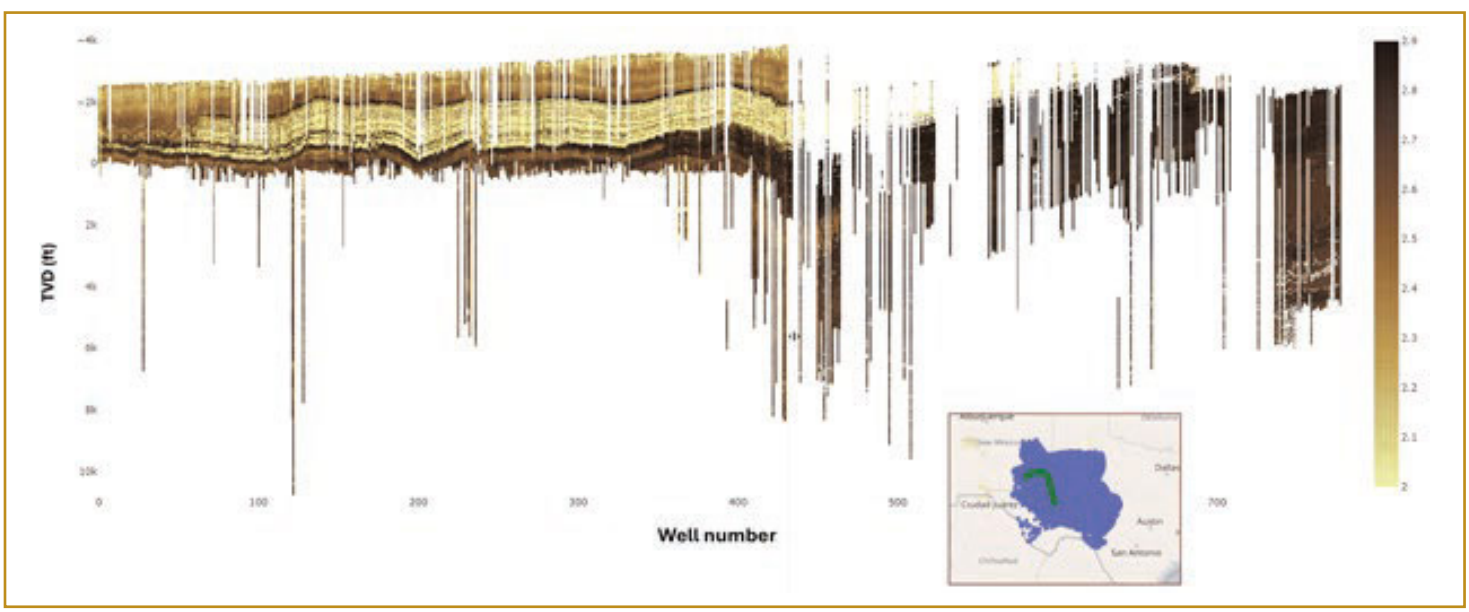

Figure 2. A cross section of well log density, the image shows a split-screen of density logs predicted by the foundation model (left), and without it (right).

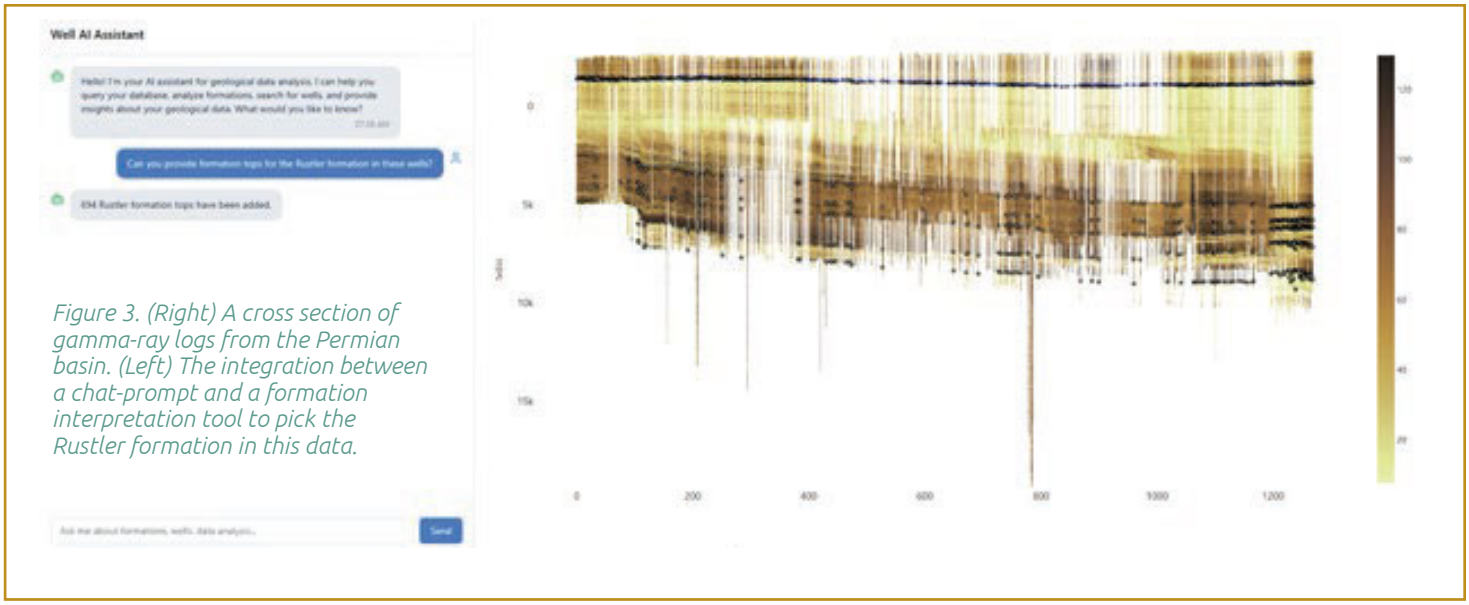

Figure 3. (Right) A cross section of gamma-ray logs from the Permian basin. (Left) The integration between a chat-prompt and a formation interpretation tool to pick the Rustler formation in this data.

◀From previous page

interpretation, TGS created a promptable formation top prediction model for the Permian Basin (figures 2 and 3). This model augments the geologist, based on their input, and can interpret in seconds what teams of interpreters would take months to do. ▶ Seismic foundation models for salt and fault interpretation: By pretraining on roughly 30 terabytes of post-stack seismic from global basins, a Seismic Foundation Model delivers state-of-the-art fault and salt delineation, as seen on figures 4 and 5. In practice, this means interpreters begin with high-quality starting points and uncertainty cues, compressing the time between survey assembly and meaningful structural interpretation. The net result is faster prospect maturation and more consistent outcomes across teams.

▶ Contracts and entitlement curation with AI agents: In today's data-driven environment, understanding your contractual rights and data license obligations is critical, but the answers are often buried across thousands of documents. Sophisticated AI agents can process these documents autonomously. By curating a knowledge base from about 140,000 documents, those agents have surfaced and linked approximately 70,000 vital documents to the specific clients and projects they govern.

This transforms the user experience. The days of "emailing legal" for routine inquiries are over. Now, any team member can use a simple search prompt to find relevant documents and interrogate them with a built-in document explorer. They receive immediate, precise answers that include the supporting clauses and a breakdown of any obligations. Crucially, the system provides the source reference, adding to "trust-but-verify."

This leads to a dramatic improvement in decision speed and quality, ensuring every choice is both fast and fully documented.

▶Quantized embeddings for retrieval at scale: Text embeddings drive relevance search across acquisition and processing archives, but full-precision vectors are expensive at scale. A seemingly modest corpus of 1,500 acquisition and processing reports, with documents ranging from 50 to 750 pages, must be segmented into semantically searchable chunks, creating a massive 250,000 embeddings, a 167-times multiplier on the original document count. This initial dataset is merely a baseline; the problem scales exponentially as the system expands to ingest the full breadth of enterprise documents, quickly transforming a

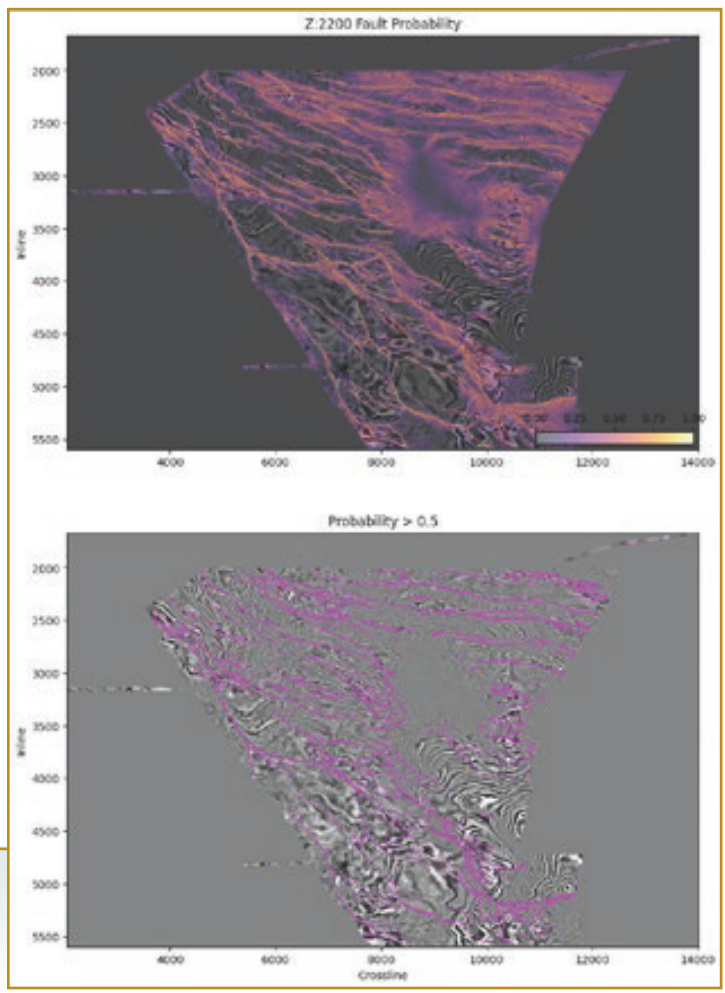

Figure 4. A time slice showing a complex of faults interpreted using a fine-tuned model. MDIO with fast access patterns makes depth slicing large datasets instant. The seismic data shown are from the publicly available Parihaka 3-D survey, provided by New Zealand Petroleum and Minerals.

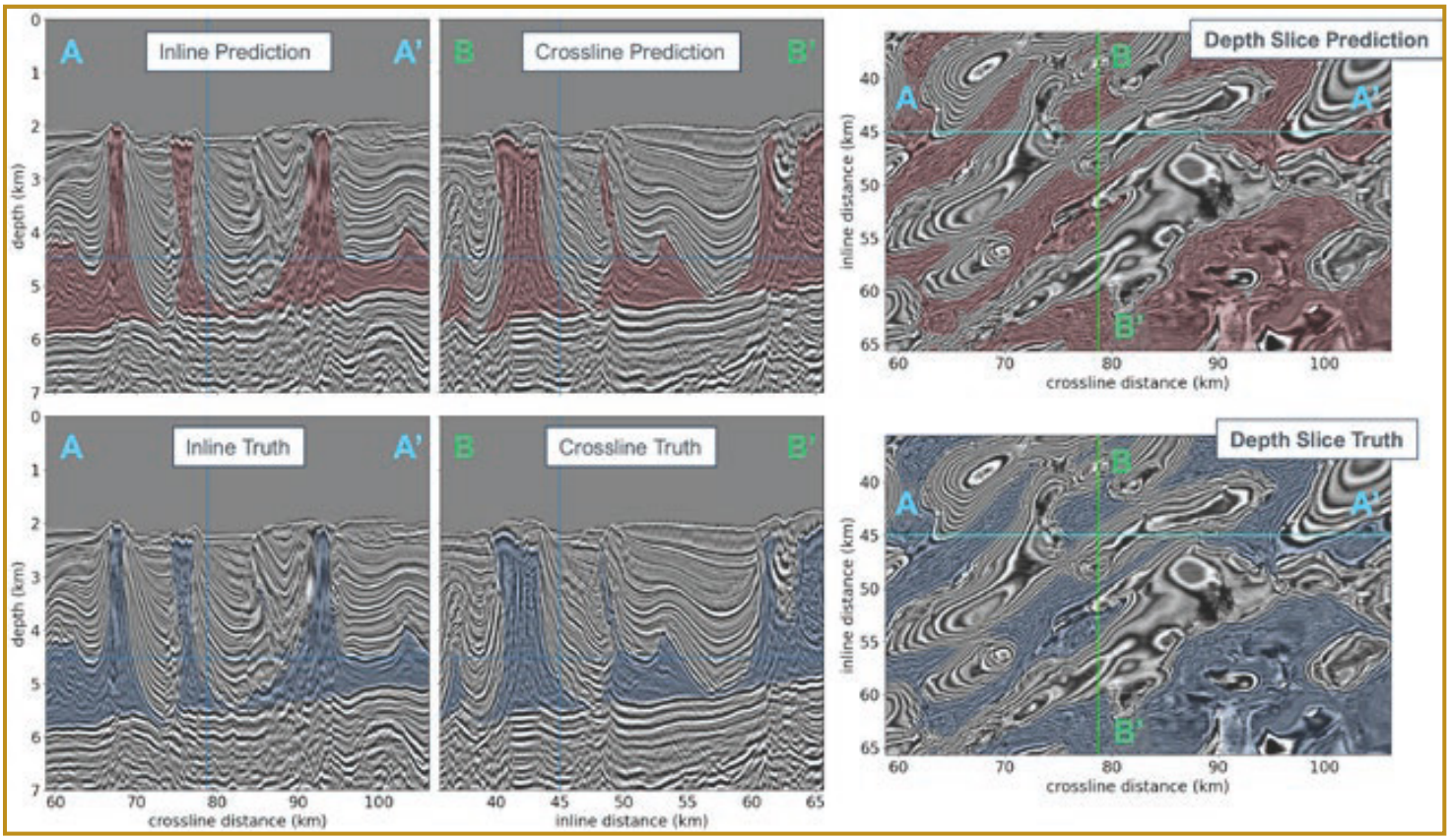

Figure 5. Automated salt interpretation trained in the Gulf of Mexico (America) and Brazil and applied in Angola. The model shows state of the art generalization "out of domain."

manageable archive into a multimillion-vector challenge that strains both budgets and infrastructure. Quantized embeddings preserve near-parity retrieval accuracy while substantially reducing storage and improving speed, enabling practical enterprise-wide semantic search without blowing the storage budget. These are not experiments for the shelf; they form the foundation of an operating model running on a configurable platform that automates routine tasks and elevates expert decision-making.

From Software to Agents

Currently organizations use a constellation of shipped software solutions, which trained experts use to bridge systems. That approach doesn't scale to today's data volume or the throughput the business expects. Agents change the shape of work. A domain-specific agent is now capable of translating user intent into action: for example, an agent might use a tool that queries SQL, performs relevance search, or pulls entitlement to data on behalf of a user. It can then inspect the results, explain the reasoning and generate an output (a dataset, a map, or a summary with citations).

Collections of agents on a shared platform, working over MDIO/Parquet stores and domain foundation models, become a living interface to the enterprise's geoscience memory. This doesn't replace software; it makes software responsive, retrieving, reasoning, verifying and documenting as it goes. Crucially, agents can be governed: they respect data entitlements, log their steps and can be audited after the fact.

What This Means for Geoscientists

For the geoscientist, it is now possible to discover all the relevant information, not just what you remember exists. Ask a question in natural language and receive a curated dataset, linked context and caveats. Routine tasks can also be automated, whether it is interpreting a seismic text header or providing high-accuracy salt and fault suggestions. Formation top picking becomes promptable. Entitlement clarity is a chat away, backed by source documents. And all this can be done faster than moving a large SEGY across an FTP server.

The outcome is not "less geoscience." It is more. The drudgery shrinks, and the work that remains is higher order: hypothesis generation, integrated interpretation, risk assessment and storytelling with evidence.

A Pragmatic Path Forward

The value of data is not in possessing it, but in processing it for purpose. TGS operates a platform on which generative AI and domain-specific foundation models interact directly with cloud-native scientific data (MDIO/Parquet) behind governed access. The agentic workflows are modular and highly configurable; most originated with little or no code via templates, with an SDK available when deeper integration is needed. Every run return source-linked outputs and an execution trace by default, so answers are inspectable. The proof points above are standard configurations, not one-off demos and the platform is designed to extend as needs evolve, adding new agents, tools and data connectors without rearchitecting.