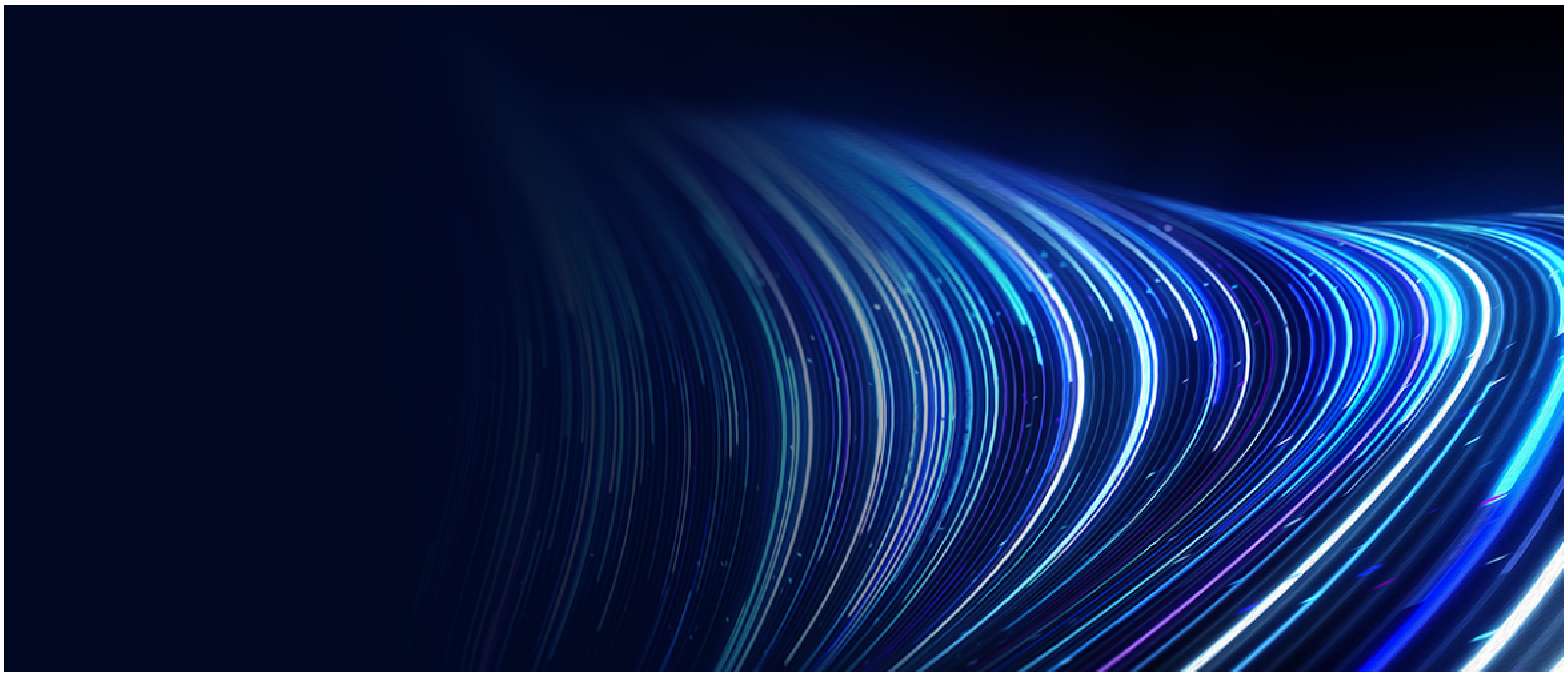
Legacy data architectures are slowing your AI journey

December 21, 2023 | Greg Thomas



What is keeping your company from capitalizing on the tremendous potential of artificial intelligence (AI) today? Your legacy data architecture is a major stumbling block.

Biopharma leaders understand that employing AI and machine learning (ML) applications can drive innovation and help transform nearly every aspect of operations—from research and development to manufacturing and quality control (QC). Biopharmas can use AI to deliver better scientific outcomes, faster, while driving down the costs of finding and producing therapeutics. Given those benefits, it's not surprising that many companies are already investing heavily in AI initiatives.

But the reality is that many companies are still years away from realizing the benefits of AI in science. Legacy data architectures—along with siloed storage environments—are significant obstacles to moving forward with AI initiatives. Until your company can implement modern, cloud-based data architectures and free your data from silos, you won't be able to use your data to generate new insights with AI.

» The problems with silos

Many biopharma organizations continue to store data in multiple isolated repositories, ranging from local hard drives and workstations to file shares and tape archives. These silos represent a decades-old approach to scientific data storage that is incompatible with advanced analytics and AI applications.

What’s wrong with silos? Moving and sharing siloed data is extremely cumbersome. In the worst case, a scientific workflow might still require using a “sneakernet”—walking a drive from one system to another, and then manually copying files onto a network file drive or an unstructured document storage environment such as Box or Egnyte.

Whenever data is distributed across the organization and maintained in legacy systems, processes are slow, collaboration is challenging, and capitalizing on AI is impossible. AI and ML algorithms require large-scale datasets associated with labeled outputs. To ensure a proper outcome, data scientists first need to understand the current data. Then they must evaluate different modeling techniques, and train and test different approaches on appropriate labeled datasets. Trying to develop and run algorithms when some of your data is stuck on disparate file shares is a Sisyphean endeavor.

Siloed environments also present serious risks. Organizations cannot implement adequate data protection and backup for every individual silo, placing that data at risk for loss. Moreover, the need to manually transfer data from one silo to another—such as an electronic lab notebook (ELN) or laboratory information management system (LIMS)—introduces the possibility of data entry errors.

To make matters worse, siloed data often remains in the proprietary formats created by scientific instrument or application vendors. Those proprietary formats lock you into small vendor ecosystems—walled gardens with limited applications where vendors are holding your data ransom. You are not able to use data in other applications or build predictive algorithms.

Before you can visualize, analyze, or use that data in AI/ML applications, you would need to prepare it for exploratory data analysis by a data scientist, then train and evaluate models. After the algorithm is developed, running an AI workload in production requires data that has been transformed into a standardized format. That format must also harmonize metadata taxonomies (definitions of data elements and structures) and ontologies (descriptions of relationships among data elements).

Data stuck in silos, locked in proprietary formats, remains static. It does not have the liquidity needed to streamline collaborative scientific work or tap into the potential of AI applications.

» Why legacy architectures can slow your journey

Some biopharma organizations have attempted to centralize data using a scientific data management system (SDMS). Traditional SDMSs were designed to store and archive data for regulatory compliance, not to prepare data for AI applications.

They might be adequate for collecting instrument and application data; cataloging data by adding some metadata; and archiving data in a compliant manner. But most traditional SDMSs have serious limitations for supporting AI initiatives.

Inflexible data flow: Traditional SDMSs have few options for data flow and processing. For example, they might be unable to send data to multiple destinations. If they can't provide the flexible data liquidity required by biopharma teams, they become a data graveyard.

Little data engineering: SDMSs are designed to store data but not transform it. Traditional SDMSs don't attempt to engineer data for scientific use cases. They don't produce data in a standardized, harmonized, future-proofed format that is engineered specifically for data science, analytics, or AI.

Poor discoverability: SDMSs might add metadata to files, but because they don't typically harmonize metadata taxonomies and ontologies, they can make it difficult for scientists to discover new or historical datasets. Data is searchable and consumable only if someone knows precisely what terms or labels to query. In many cases, lab scientists end up re-running an assay or an experiment because that's easier than finding historical data.

Inflexible accessibility: SDMSs are certainly several steps above thumb drives. But they are still largely closed, siloed data repositories. Traditional SDMSs require users to access data only through the SDMS interface, not through their usual interfaces and applications, such as ELNs, analytics tools, or AI applications.

Lack of scalability: On-premises SDMSs cannot be scaled easily or cost effectively: Each upgrade requires multiple changes, including upgrades for the database, servers, and file storage. If SDMSs employ cloud services at all, they often use the cloud as another data center. Consequently, SDMSs are not the best environment for assembling the large-scale datasets required for AI.

SDMSs simply aren't designed to prepare data for AI. Some SDMS vendors might tack on capabilities to address deficiencies. But in general these legacy solutions cannot provide sufficient data liquidity, allow adequate searchability, enable data accessibility, or efficiently scale up to support the massive volumes of AI-native data needed for AI algorithms.

(Read more about why a traditional SDMS is not an option anymore.)

» Before you can close the AI gap

When it comes to AI, there is a large gap between the goals of biopharma executives and the reality in labs. Until biopharma organizations can address key data obstacles, they will be unable to realize the benefits that AI can deliver for science. Retaining a legacy data architecture in the form of a traditional SDMS, and leaving data in siloed environments, prevent organizations from producing the open, vendor-agnostic, purpose-engineered, liquid, large-scale data that they need for AI applications.

Unfortunately, legacy architectures and data silos are not the only obstacles. A do-it-yourself (DIY) approach to data management plus a lack of data standardization and harmonization can also slow your progress. Sufficiently addressing all of these obstacles will be necessary before you can accelerate your AI journey and improve scientific outcomes.

Learn more about the AI gap in the white paper, “The Scientific AI Gap.”