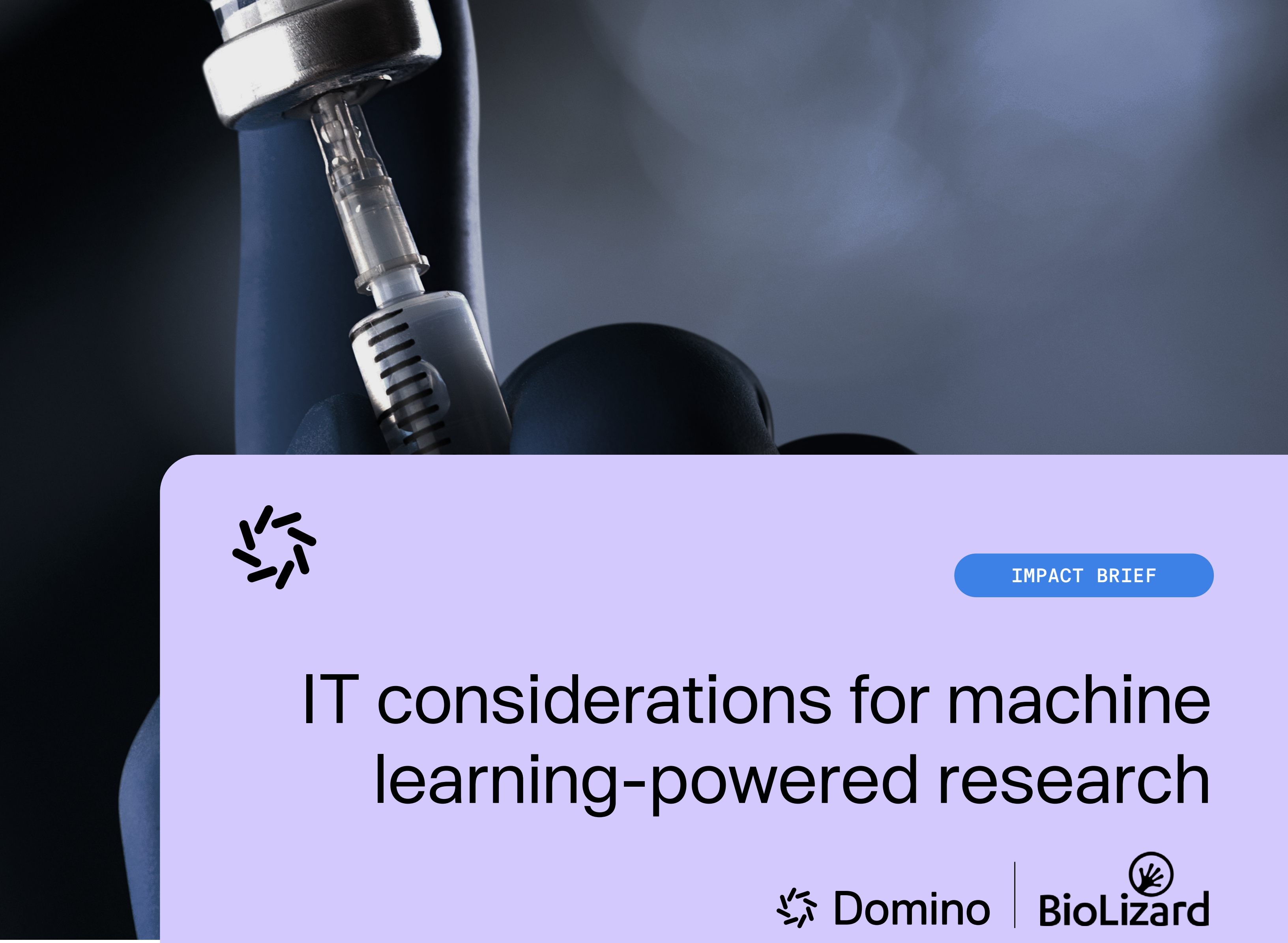

IMPACT BRIEF

IT considerations for machine learning-powered research

Domino | BioLizard

Bridging the gap between data management, analytics, and MLOps for bioinformatics

The era of big data, artificial intelligence, and machine learning (AI/ML) requires scalable computing resources and data platforms to: 1) efficiently ingest, process, and analyze the massive amount of data generated during research workflows, and 2) machine learning systems that empower the organization to leverage and explore new research opportunities. Versioning and encapsulating code in containers while maintaining workflows and pipelines are some of the new strategies needed to extract value from qualitative scientific results. As such, IT architectural considerations must balance best practices across data management/governance and DevOps/MLOps — a common challenge faced by many IT leaders assessing data/analytics and machine learning platforms. Deploying production models requires a factory-like approach using a structured framework to promote the reuse, governance, and maintenance of models.

Generating scientific results that can be assessed, verified, and reproduced is a critical objective in pharmaceutical research. While reproducibility has always been a core aspect of research, innovations across data management, data science, machine learning, and AI underscore the need for sustainable, maintainable, scalable, and reproducible practices. Processes spanning multiple research teams using different infrastructures and workflows create siloes slowing the delivery and investigation of new research avenues.

Critical IT evaluation criteria for bioinformatics and data/analytics platforms

Machine learning platforms ensure consistent research pipelines

Robust data and machine learning systems should enable the entire organization to leverage and explore new research opportunities to create value — while maintaining consistent pipelines for governance and maintenance. For robust data analyses and machine learning, essential functionalities include interactive notebooks or IDEs, shell/terminal access, and environment management. Workflow execution is further enhanced by integrating external libraries, managed version control, container repositories, support for modular code component use/reuse, integration with CI/CD pipelines, profiling/complexity monitoring, and automated testing via scripts.

The machine learning lifecycle, including model development and training, requires a platform that supports distributed compute clusters, hyperparameter optimization, MLOps workflow/tooling integration, AutoML support, and synthetic data generation. Data science teams need experiment tracking and metadata storage for full reproducibility, as well as pre-built models and projects to accelerate the modeling process. Finally, model lifecycle management ensures consistent performance and smooth deployment.

Data lineage and auditability for governance, traceability, and reproducibility

When selecting data and analytics platforms, IT leaders must consider several key functionalities when it comes to bioinformatics. Out-of-the-box data connectors for integrating common data sources (both internal and external) should support both retrieval and manipulation, particularly for structured data. A sandbox environment is essential for safe data manipulation without impacting live systems. Visual tools for automated data transformations and automated quality control (QC) maintain data accuracy and consistency across data and analytics teams. Additionally, data versioning and privileged access enhance security, while automated lifecycle management optimizes resource handling. Data lineage and auditable logs provide traceability and reproducibility for governance, and data labeling improves usability for machine learning applications. Testing datasets is crucial for validating the accuracy and functionality of data management processes, systems, and tools.

Workflow orchestration and automation across multiple teams and users

For efficient workflows in bioinformatics, key requirements include a workflow GUI builder for designing pipelines without coding and workflow orchestration to coordinate multiple tasks. Support for Nextflow to run data-driven applications on distributed computing machines, containers, and scalable infrastructure are essential for packaging and deploying software applications and dependencies. Critical features should include alerting on insufficient resources, reentrancy for resuming interrupted processes, and reproducibility to ensure consistent results.

Further, testing support for applications and pipelines, scripted setup for automating dependencies, and interactive workflows to monitor workflow/job status greatly enhance functionality. Additionally, awareness and tooling to handle chemical/biological data, trigger detection of environment or data changes, resource monitoring with auto-healing to reduce bottlenecks, and comprehensive workflow/job tracking ensure robust and responsive workflow management for bioinformatics and IT leaders.

Bioinformatics and data/analytics platforms must enable research teams to **drive value**

IMPACT BRIEF

About BioLizard

BioLizard is a leading multinational data analytics, AI, and data management consulting company supporting digital transformation in the life sciences industry. By offering custom solutions in data analytics, bioinformatics, software development, AI, and more, we support biopharma and biotech companies in accessing vital data-driven insights to drive their life sciences research forward.

As well as providing custom solutions ourselves, we also provide unbiased vendor assessments for clients, to identify the data science solution providers that best fit their unique needs and goals.

About Domino Data Lab

Domino Data Lab empowers the largest AI-driven enterprises to build and operate AI at scale. Domino's Enterprise AI Platform unifies the flexibility AI teams want with the visibility and control the enterprise requires. Domino enables a repeatable and agile ML lifecycle for faster, responsible AI impact with lower costs. With Domino, global enterprises can develop better medicines, grow more productive crops, develop more competitive products, and more. Founded in 2013, Domino is backed by Sequoia Capital, Coatue Management, NVIDIA, Snowflake, and other leading investors.

Learn more at **www.domino.ai** →

Domino

From prototype to production: Key IT considerations across the model lifecycle

Key value considerations

Overall, bioinformatics and data/analytics platforms must enable research teams to drive value. Collaboration and continuous improvement provide significantly more value to researchers through efficient and transparent workflow/processes, knowledge/documentation management, improved reproducibility and pattern recognition, closing knowledge loops with clear delineations across data platform usage and access. Teams ultimately should be able to efficiently test hypotheses with accurate and efficient results.

Bridging the gap from prototype to production

For effective model deployment and serving, key platform functionalities include the ability to annotate datasets for ML and analyses, create interactive visualizations and dashboards using model output, and explainability of model predictions or outcomes. Interoperability is key — data or analysis results should be exportable to other tools or formats and integrated with various data sources such as APIs (particularly REST APIs), files, dashboards/applications (including Shiny Server connections), and events in data platforms. Finally, templatized analysis support with pre-built templates/frameworks, automated deployment and rollback, and real-time performance monitoring for correctness, performance, cost, and other metrics is essential.

Visit domino.ai/partners/biolizard →