How Elucidata Enables AI-Ready Collaboration Across Institutions: Securely and at Scale

Authored by:

Harshveer Singh, Director of Engineering, Elucidata

Swastik Gowda, Senior Software Engineer, Elucidata

WHITEPAPER

elucidata

Contents

The Authors

Harshveer Singh, Director of Engineering, Elucidata leads cross-functional product teams across front-end, back-end, and cloud infrastructure, while overseeing security and program management. With over a decade of experience, he brings deep expertise in Python, Node.js, Angular, and AWS. He builds scalable, cloud-native systems, including microservices and serverless platforms, for complex, data-intensive applications in the life sciences.

Swastik Gowda, Senior Software Engineer, Elucidata, designs and builds scalable, cloud-native infrastructure for biomedical data processing and AI-driven applications. He leads initiatives across Kubernetes, AWS, and independent cloud platforms, focusing on cost optimization, high-performance computing, and secure data workflows. Swastik has driven key cost-optimization efforts, achieving over 80% savings. He also plays a pivotal role in translating HIPAA compliance requirements into robust, secure engineering solutions.

Executive Summary

Healthcare AI is stalling! Not due to a lack of data, but because valuable multimodal data (omics, imaging, clinical) is trapped in institutional silos. Centralized ML approaches hit walls: regulatory hurdles, privacy risks, and subpar generalizability. But the real barrier? Making biological data usable for AI through curation, normalization, and context-aware feature engineering without violating compliance or burning resources.

Elucidata's Federated Learning Solution rewrites this equation. Our approach allows institutions to train models locally without ever moving raw data while tapping into Elucidata's domain-specific preprocessing pipeline that harmonizes and enriches data at each node. This ensures each site contributes high-quality, AI-ready inputs, with zero compromise on privacy.

In a recent deployment, our platform enabled a model to **predict gene expression from whole-slide pathology images across three sites**. No raw data was moved, training happened securely across both cloud and on-prem environments. Thanks to Polly, our AI-ready data platform with built-in observability, teams saw 70% faster iteration, real-time monitoring, and transparent model evaluation.

The result?

Better generalizability across institutions

Lower infrastructure and compliance costs

Real-world readiness for cross-institutional AI collaboration

This is what **scalable, secure, and scientifically grounded AI** looks like.

Introduction

Imagine having access to a treasure trove of biomedical data across dozens of leading institutions, genomic sequences, high-resolution pathology images, and comprehensive clinical records, yet not being able to harness this wealth for AI model development due to privacy regulations and institutional silos. This scenario isn't hypothetical;

IN

97%

of healthcare organizations valuable data remains locked away.

This results in subpar AI models trained on limited, homogeneous datasets that fail to generalize across diverse patient populations.

It then becomes important to discover ways how qualitative AI models can be generated and trained to do away with the limitations of working with subpar models. In this context, Federated Learning Solution has recently emerged as a viable solution, which is transformative in nature.

Background and Context

The global healthcare AI market is projected to exceed $188 billion by 2030 (Precedence Research, 2024), driven by demand for precision diagnostics and accelerated drug development. Yet, its progress hinges on overcoming a critical barrier namely unstructured and siloed data.

In the digital economy, there is no dearth of data but it being fragmented and devoid of a common access point, often causes delay in the process of R&D, and thus drug development.

While the availability of this humongous sensitive data in healthcare is certainly beneficial for training ML models, rising concerns about data privacy and security have made centralized storage quite difficult. Federated Learning (FL)-an approach in which training data is not managed centrally, has acquired significance in terms of its peculiar capacity to allay privacy concerns and ensure security.

The healthcare industry with its voluminous data in the form of EHR, genomic sequence and diagnostic imaging carries immense potential for transformative solutions, when utilized effectively.

In this context, Elucidata's Federated Learning solution is uniquely positioned, and enables access to siloed and sensitive healthcare data in a secure and compliant way.

Our solution promises to be of immense value to healthcare organizations across the globe as it meets all the significant privacy and security compliance parameters, and address the following issues effectually:

01 **Regulatory Pressure:** GDPR, HIPAA, and emerging frameworks (e.g., EU AI) Act: penalizes non-compliant data practices.

02 **Data Complexity:** Multimodal datasets (e.g., spatial transcriptomics, digital pathology) require domain-specific preprocessing to avoid "garbage-in, garbage-out" AI.

03 **Collaborative Imperative:** Landmark initiatives like the EU's 1+ Million Genomes Project demand privacy-preserving collaboration across borders.

Significance

Federated Learning is revolutionizing healthcare by enabling AI models to be trained across multiple institutions without sharing sensitive patient data. Its significance lies in key features:

Privacy First AI

Enhanced accuracy

Faster innovation

Models trained on homogeneous data

- Struggle to perform effectively when applied to diverse, real-world data.
- Aggregating sensitive datasets- risk breaches, ownership disputes, and models biased toward single-institution data.

Models trained on diverse data

- Perform much better in the real world.
- It highlights the need and urgency to adopt Federated Learning approach to stay competitive in the market by training models on geographically and ethnically heterogeneous patient data.

Objectives

This paper aims to:

01 **Introduce Elucidata's Federated Learning solution**, explain how it is different from generic solutions that already exist, and what are the strategic and operational benefits of using this solution.

02 Showcase how this solution will help in adherence to regulatory compliances, reduce overall training time with built-in observability, leading to **accelerated drug discovery**.

03 **Elaborate on the preprocessing of raw data** by domain experts as a part of Elucidata's FL solutions and how that will help in producing high-accuracy model predictions.

Problem Statement

The healthcare industry faces a fundamental issue: while the potential for AI-driven breakthroughs depends on access to large, diverse datasets, patient data remains fragmented across institutions due to privacy regulations, competitive concerns, and technical barriers. These data silos severely constrains the development of robust machine learning models that could otherwise revolutionize patient care, drug discovery, and clinical research.

Critical Challenge

Developing mechanisms to make siloed healthcare data accessible for collaborative research and model training without compromising patient privacy, institutional autonomy, or regulatory compliance requirements such as HIPAA, GDPR, and emerging data governance frameworks.

Challenges and Impact of Siloed Data

Healthcare data remains trapped within individual institutions due to competitive dynamics and regulatory constraints, which prevents the formation of comprehensive datasets needed for robust AI model development.

Impact

- **Scientific:** Siloed data often represents a limited subset of the population or research conditions, leading to datasets that are unrepresentative of broader diversity. AI models trained on such data are prone to skewed predictions, disproportionately favoring specific groups or outcomes.

- **Operational:** Individual institutions must invest abundantly in AI infrastructure and isolated AI initiatives that could benefit from shared resources.

Limitations of Centralized Machine Learning in Healthcare

Centralized ML requires aggregating terabytes of sensitive healthcare data (imaging, genomics, EHRs) into a single repository. Centralized repositories create attractive targets for cyberattacks and represent catastrophic failure points where a single breach can compromise millions of patient records. Also, the institution and infrastructure when data is being aggregated need to adhere to various regulatory compliances while handling these datasets.

Impact

- **Scientific:** In centralized ML, due to various data aggregation limits, data freshness is an issue. Sometimes, data might be 6-18 months old, which reduces model relevance. Also, breaches can halt a project for months, causing unnecessary delays.

- **Technical:** The UK Biobank project, despite its success, required over a decade and hundreds of millions in funding primarily due to data centralization challenges across the NHS system.

1. Privacy and Compliance Risks

Centralizing sensitive patient data (e.g., genomic records, WSIs) exposes institutions to breaches and regulatory penalties. This is mainly due to Data residency laws (GDPR Article 44), evolving consent frameworks, and increasing cyberattack sophistication (e.g., ransomware targeting healthcare).

Impact

- **Scientific:** Restricts access to diverse datasets, limiting model generalizability.

- **Operational:** Breach mitigation costs healthcare organizations **$10.1M** on average (**IBM 2023**).

- **Regulatory:** Non-compliance fines (e.g., €20M under GDPR) deter collaboration.

2. Data Ownership and Trust Barriers

Institutions resist sharing data due to IP concerns and mistrust in third-party platforms, stemming from the lack of standardized data-sharing agreements and transparency in model and data usage.

Impact

- **Scientific:** Siloed datasets produce models biased toward local populations (e.g., 80% of genomic data is from European ancestry (Li et al., 2020).
- **Operational:** Negotiating data-sharing terms delays projects by 6–12 months.

3. Technical and Operational Fragmentation

Different data modalities (Imaging, Omics, EMR etc.), legacy systems, and variable compute resources hinder cross-institutional workflows due to decentralized IT procurement and a lack of interoperable standards.

Impact

- **40% of AI projects fail** during data integration **(Gartner, 2023).**
- Manual data harmonization **consumes 70% of data scientists' time (Press, 2016).**

4. Limited Model Generalizability

Models trained on narrow datasets underperform in the real world because Centralized training data lacks geographic, demographic, and technical diversity.

Impact

- Studies have shown that **10-20% AUC (area under the curve) decreases when sepsis models are deployed at external sites (Sendak et al., 2020)**

These barriers frequently cause healthcare AI projects to stall or underperform. As healthcare data becomes more complex, the need for collaborative and privacy-preserving approaches grows even more critical.

Elucidata's Federated Learning Solution

A Foundation for Privacy-Preserving AI in Healthcare

Using Elucidata's Federated Learning solution, each participant, such as a hospital or research lab, will train a local version of a shared model using only its private data.

Instead of sharing raw data, they send updated model parameters (like weights or gradients) to a central server, which aggregates these parameters to improve the global model. Before training, Bioinformaticians will harmonize the data (e.g., batch correction for transcriptomics, WSI stain normalization). This process iteratively refines the model, without any raw data ever leaving its source.

Elucidata's FL Solutions: Transformative for Healthcare Sector

Elucidata's Federated Learning Solution can be transformative for healthcare owing to following attributes:

Enhanced Data Privacy & Security

Raw patient data remains secure on-site at each institution, minimizing the risk of unauthorized exposure while ensuring compliance with patient consent agreements and privacy regulations.

Collaborative Innovation Without Compromise

Healthcare institutions can work together to build robust, generalizable AI models that benefit from diverse patient populations and clinical practices, all while keeping sensitive data within their own secure environments.

The foundation of successful federated learning in healthcare rests on trust and regulatory alignment. Institutions maintain complete ownership and control over their data assets, while our platform, Polly, provides comprehensive audit logs that track model update provenance for full transparency.

By keeping data local and eliminating cross-border transfers, this approach naturally supports compliance with stringent healthcare regulations, including HIPAA, GDPR, and other regional privacy laws. This in turn, effectively removes the legal barriers and data residency challenges that traditionally complicate multi-institutional research and collaboration.

Real-World Relevance

Recent market intelligence underlines the urgency and opportunity. According to Precedence Research, the global AI in healthcare market could reach over $600 billion by 2034 (Precedence Research, 2025), but only if regulatory, privacy, and scalability barriers are overcome. Federated learning is poised to unlock this latent value while enabling data collaboration and aligning with data governance in healthcare and life sciences.

Uniqueness of Elucidata's solution

Biological Data Mastery

Elucidata provides specialized pipelines and models for biomedical data processing for varied ML applications.

Polly Observability

Real-time tracking of data drift (e.g., batch effects in WSIs) and model fairness metrics.

Compliance-by-Design

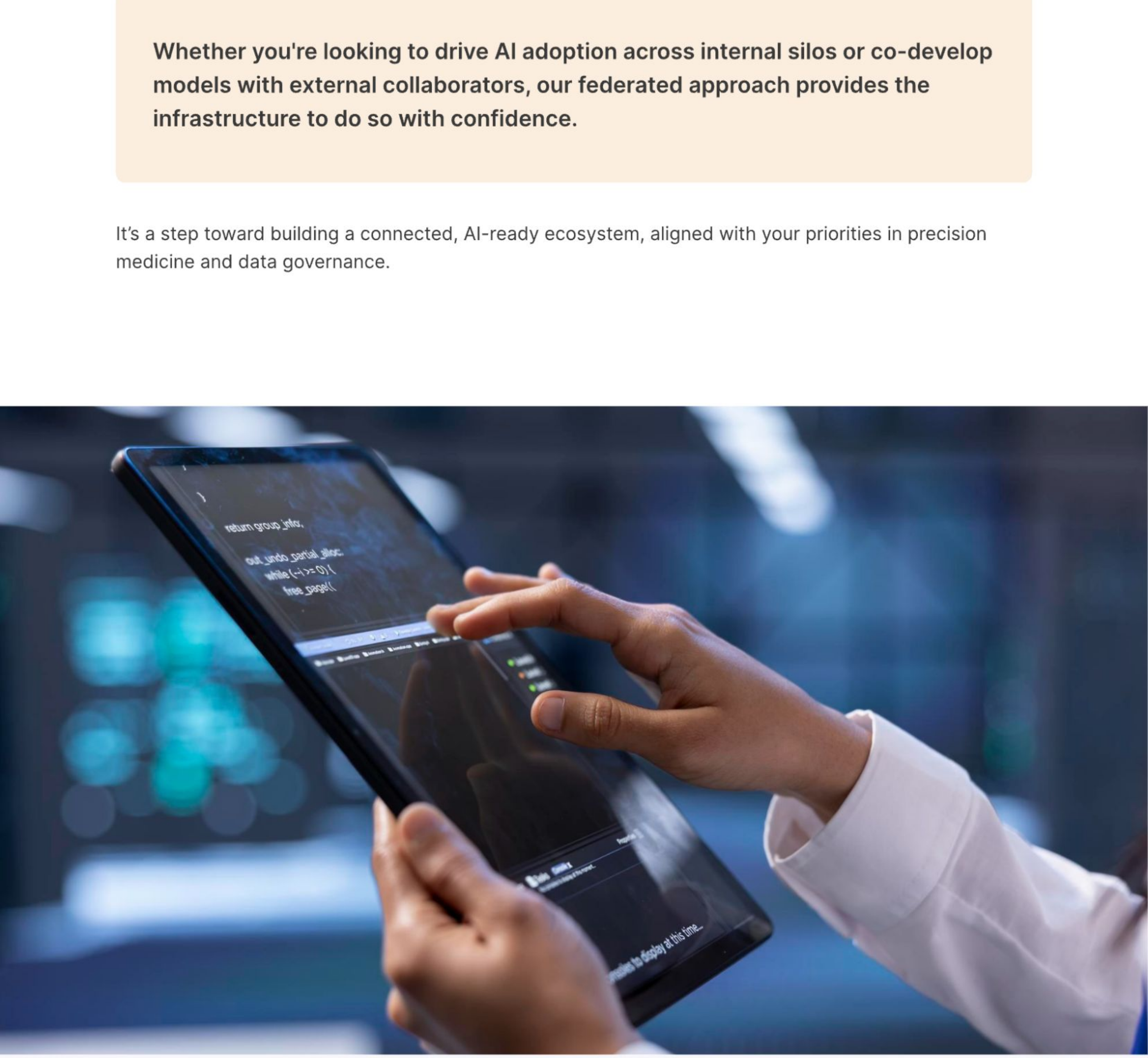
Preconfigured workflows for HIPAA/GDPR, including automated Data Protection Impact Assessments (DPIAs).

Strategic Alignment

Elucidata's Federated Learning solution is designed to support your organization's need for secure, collaborative AI, particularly when working across institutional boundaries with partners, CROs, or hospital systems. As an extension of Polly, our data-centric AI platform, this solution enables model development without exposing sensitive data, helping your teams accelerate innovation while staying compliant with evolving privacy and regulatory standards.

Whether you're looking to drive AI adoption across internal silos or co-develop models with external collaborators, our federated approach provides the infrastructure to do so with confidence.

It's a step toward building a connected, AI-ready ecosystem, aligned with your priorities in precision medicine and data governance.



Elucidata's Federated Learning Architecture

Secure, Compliant Collaboration at Scale

Elucidata's federated learning solution is built on AWS services, utilizing managed compute, storage, and networking solutions as shown in Figure 1. On top of this, we have specialized libraries and implementation procedures to quickly setup FL infrastructure which adheres to security standards.

This solution also demonstrates how FL can be made practical, secure, and production-ready for ML model development, such as gene expression prediction. The recent case study at Elucidata, implemented this architectural setup, having 3 participant nodes and one central aggregation server.

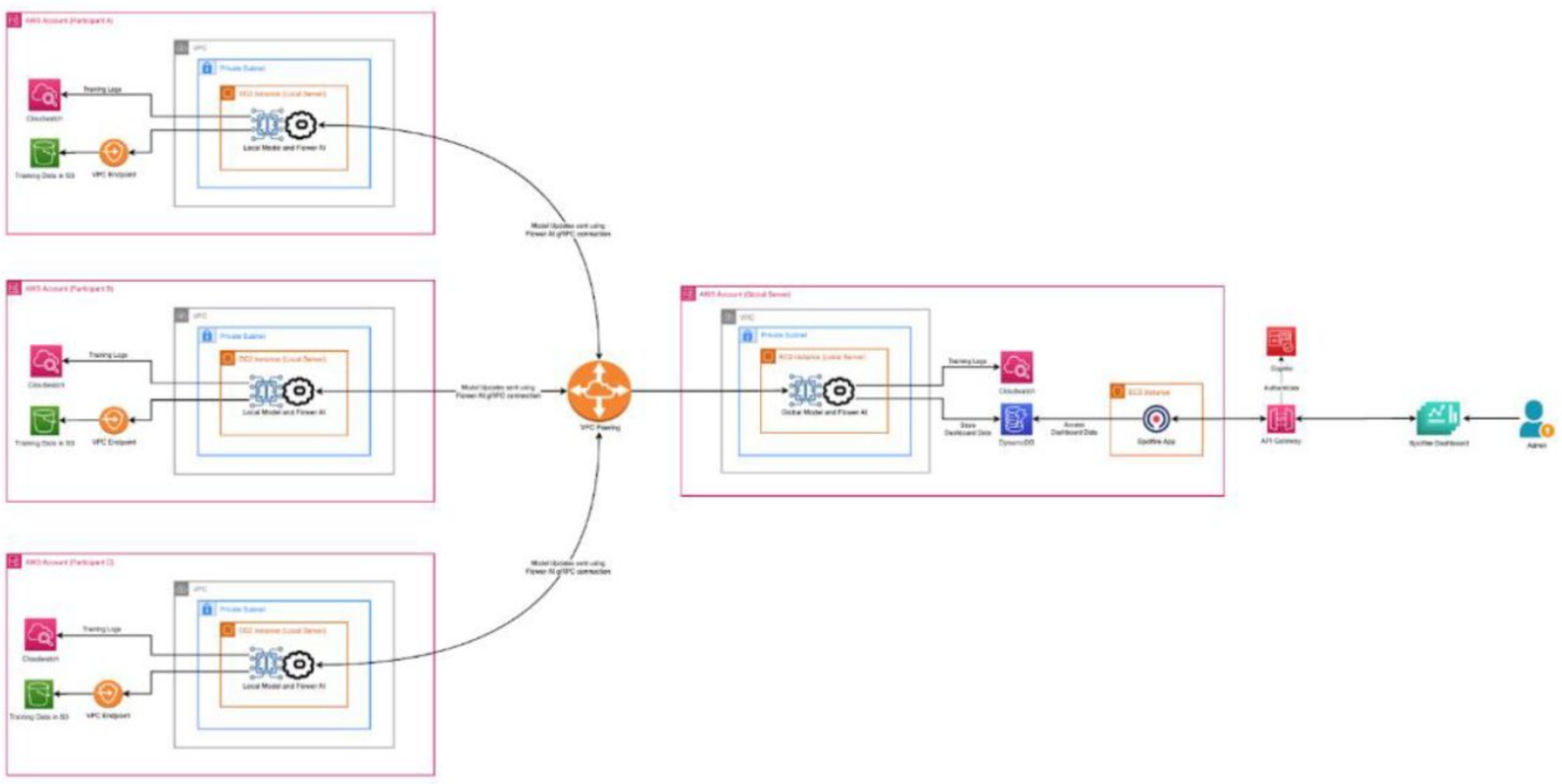


Figure 1: Architecture implemented in this Case Study

System Design and Infrastructure

On a high level, the system comprises three primary components: distributed participant nodes, a central aggregation server, and the Polly observability layer.

Participant Node Architecture

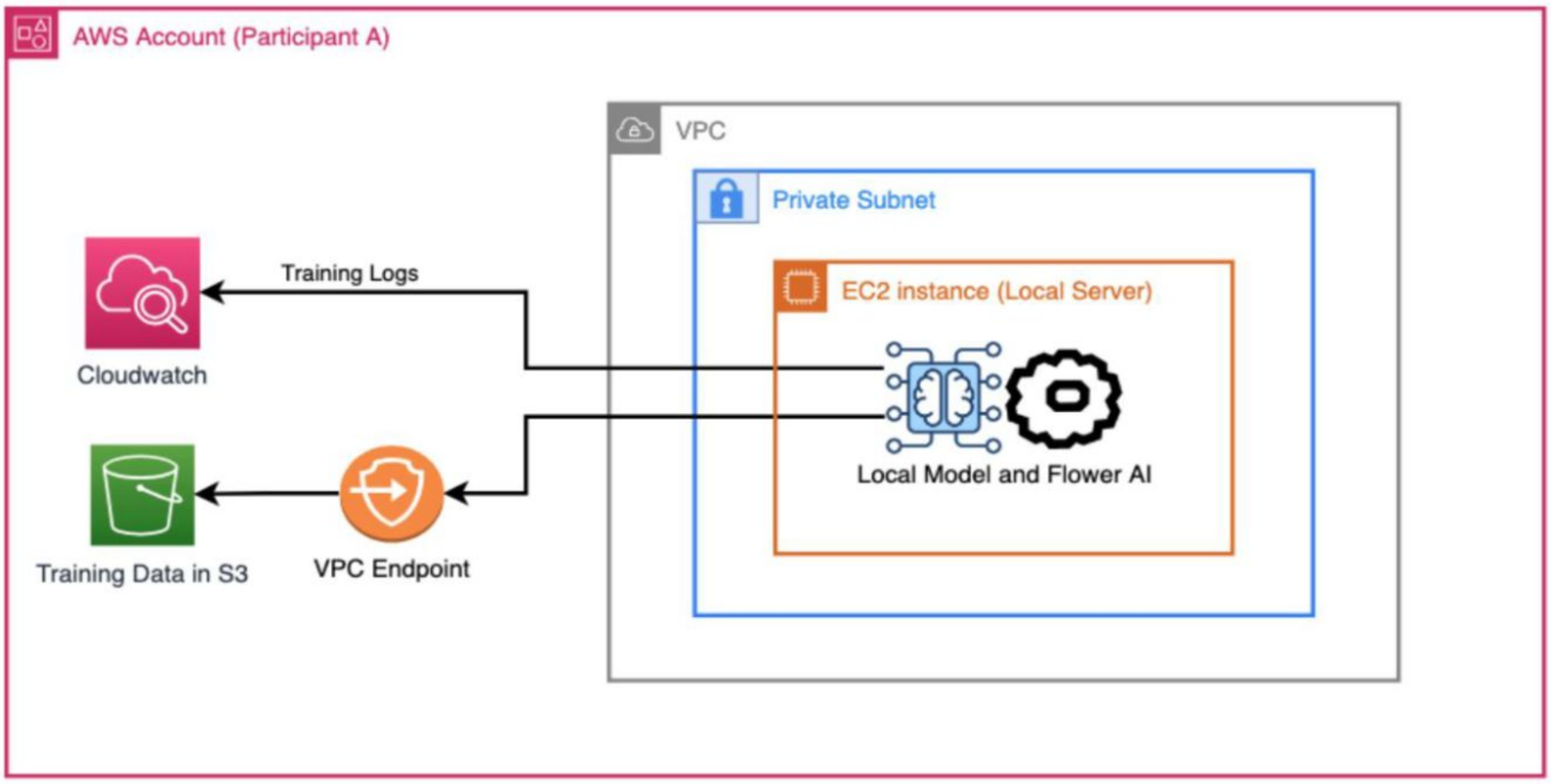


Figure 2: Participant Node

Each institution participates as an independent entity, hosting its private data on its own secure AWS environment. The solution described orchestrates local model training across three AWS accounts acting as participants, each with:

- **Secure Data Storage (Amazon S3):** S3 buckets with server-side encryption and versioning enabled, accessed exclusively through VPC Gateway Endpoints.
- **Compute Environment (AWS EC2):** Auto-scaling EC2 instances (g4dn.12xlarge) deployed within private subnets, ensuring complete network isolation.
- **Security Perimeter:** Network ACLs and security groups implementing defense-in-depth, with all ingress from the public internet blocked.

Central Aggregation Infrastructure

The aggregation server, hosted in a dedicated AWS account, orchestrates the federated learning process. After each round of local training, only model parameters (not raw data) are pushed to a central server.

- **Flower Framework Server:** Coordinates training rounds, manages client connections, and implements FedAvg algorithms.
- **Parameter Storage:** Encrypted S3 buckets store model checkpoints and aggregated weights with lifecycle policies for version management.
- **Communication Layer:** VPC peering connections with each participant.

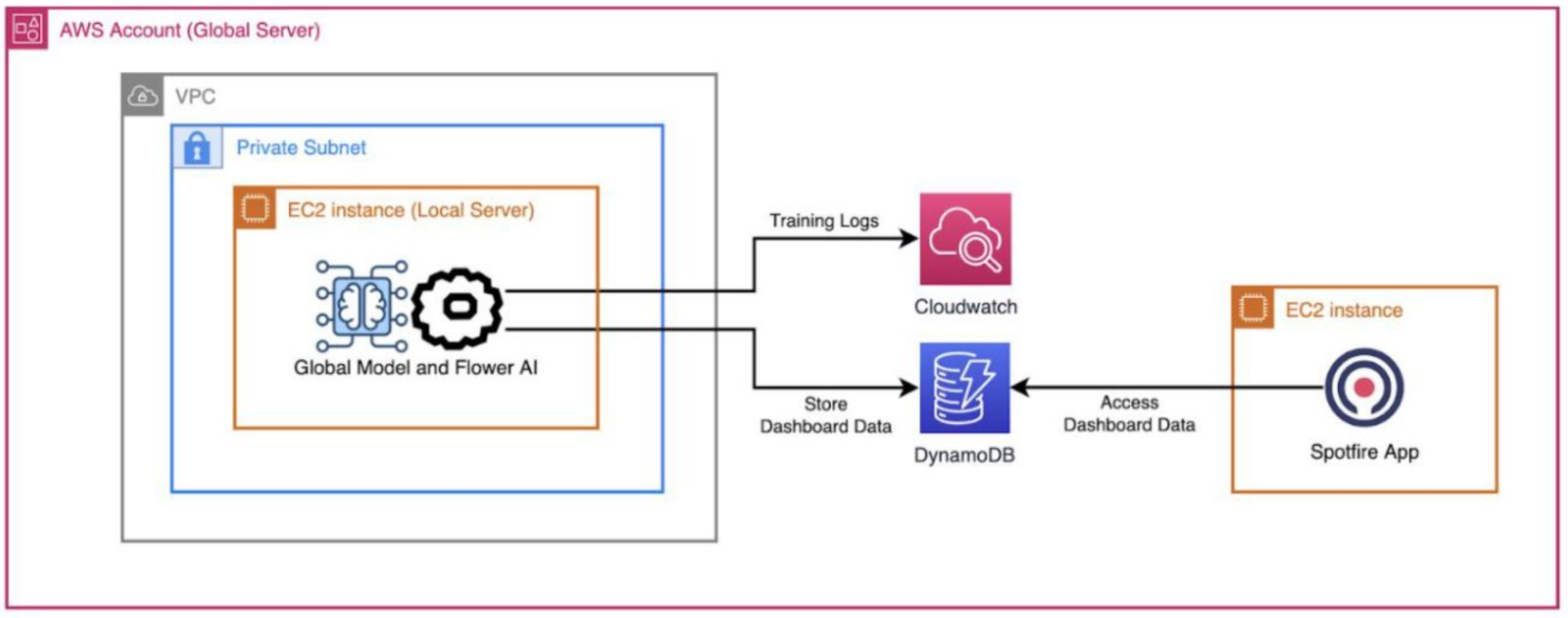


Figure 3: Central Aggregation Infrastructure

Security and Privacy Architecture

Our zero-trust security model implements multiple layers of protection:

- **Differential Privacy Engine:** Using Flower, institutions can tune their differential-privacy settings by clipping model updates and adding Gaussian noise to the aggregated model.
- **IAM Policy Framework:** Least-privilege access enforced through role-based policies, with cross-account assume-role patterns for federation.
- **Audit Infrastructure:** CloudWatch Logs aggregation with real-time anomaly detection and compliance reporting.

Scalability and Interoperability Design

The architecture supports horizontal scaling through:

- **Dynamic Participant Registration:** New institutions join via automated CloudFormation templates, reducing onboarding from weeks to hours.
- **Protocol Agnostic Communication:** gRPC-based messaging supports heterogeneous compute environments (on-premise, multi-cloud).
- **Resource Elasticity:** Auto-scaling groups adjust compute capacity based on training workload, optimizing cost-efficiency.

Monitoring, Audit, and Visualization

Comprehensive observability across the federated learning infrastructure ensures operational transparency, regulatory compliance, and performance optimization throughout the distributed training process.

The monitoring stack provides real-time insights into both technical metrics and business outcomes while maintaining strict audit trails for compliance requirements.

- **Polly Dashboards** provide centralized dashboards for monitoring and visualization of ML training loss, accuracy, and epoch.
- **Amazon CloudWatch** captures comprehensive logs for audit and troubleshooting.

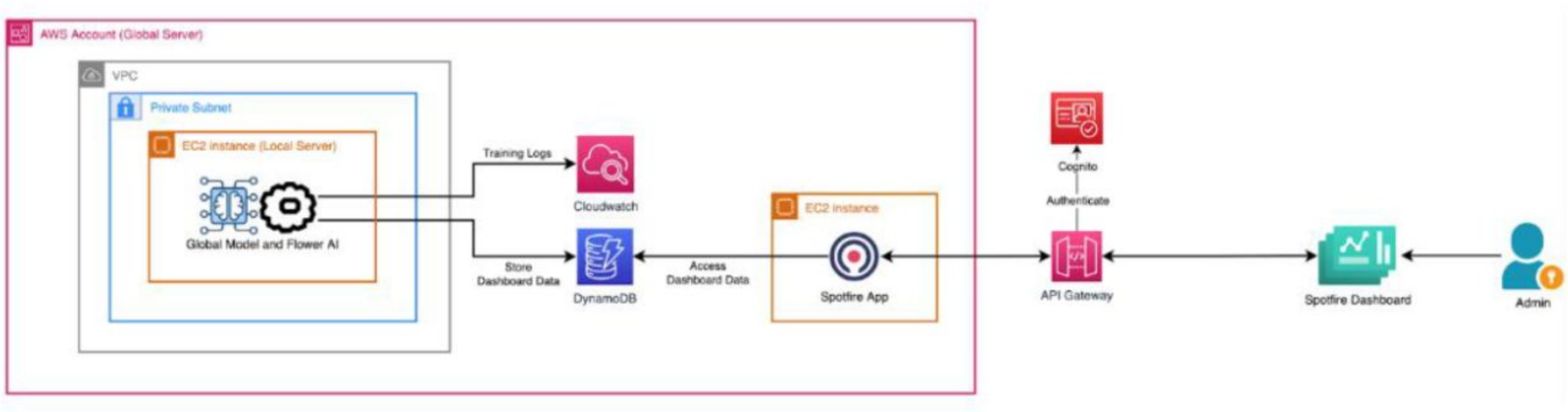


Figure 4: Central aggregation with logs and visualization

Key Components and Security Functions

The table below outlines the key infrastructure elements and security controls that underpin Elucidata's privacy-preserving federated learning solution.

Component	Function
AWS EC2	Local model training, within the private network
Private Subnet/No Internet	Further isolates the training environment
Amazon S3 + S3 Endpoint	Secure data storage, AWS private backbone
VPC Peering	AWS Secure, private communication backbone
Global Server	Parameter aggregation & model update
IAM & Encryption	Least privilege and data confidentiality
Amazon CloudWatch	Logging, monitoring, and alerting
Flower FL Framework	Local/global cooperation orchestration
Polly Dashboard	Performance and audit visualization

Case Study

Federated Gene Expression Prediction from Histopathology Images

This case study presents a comprehensive implementation of federated learning in precision oncology. We will discuss technical implementation, results, and the teams involved in this case study to demonstrate the application of Elucidata's federated learning solution.

Project Overview

Elucidata demonstrated the viability of federated learning for precision oncology by training a deep learning model to predict gene expression profiles directly from WSI images across three geographically distributed simulated institutes.

Participants

The federated gene expression prediction project required a multidisciplinary team with expertise spanning machine learning, bioinformatics, pathology, and cybersecurity. This diverse collaboration was essential to address the complex technical and regulatory challenges inherent in medical AI applications. The team structure ensured that both the scientific rigor and security compliance standards were maintained throughout the project lifecycle. Therefore, the following teams were constituted to do this case study.

- **Technical Team:** 2 ML engineers, 1 bioinformatician, 1 expert on pathologic images, 1 security compliance officer.
- **Infrastructure Providers:** AWS professional services for architecture review and optimization.

Technical Implementation

The implementation leveraged our comprehensive federated learning infrastructure to simulate a realistic multi-institutional collaboration scenario. We created an authentic representation of the geographical and network constraints that would exist in real-world federated learning deployments by deploying across three distinct AWS regions.

- **Model Architecture:** HE2RNA, A deep-learning algorithm specifically customized for the prediction of gene expression from WSI (Nature Communications, 2020).
- **Training Infrastructure:** 3 AWS accounts with different regions, simulating geographic distribution.
  - Client 1: 580 WSIs
  - Client 2: 290 WSIs
  - Client 3: 289 WSIs
- **Data Volume:** 800 GB
- **Total number of Whole Slide Images (WSIs):** 1159 WSIs
  - Client 1: 580 WSIs
  - Client 2: 290 WSIs
  - Client 3: 289 WSIs
- **Federated Protocol:** FedAvg with momentum, 4 rounds, 120 epochs per round.

Results and Performance Metrics

The federated learning approach successfully demonstrated both technical feasibility and privacy preservation. The model achieved meaningful predictive performance while maintaining strict data privacy standards throughout the training process.

- **Model Performance:** Achieved 0.246 Pearson's correlation between ground truth and predicted gene expression
- **Privacy Preservation:** Zero data exposure events
- Visualization of Train Loss Value vs Step

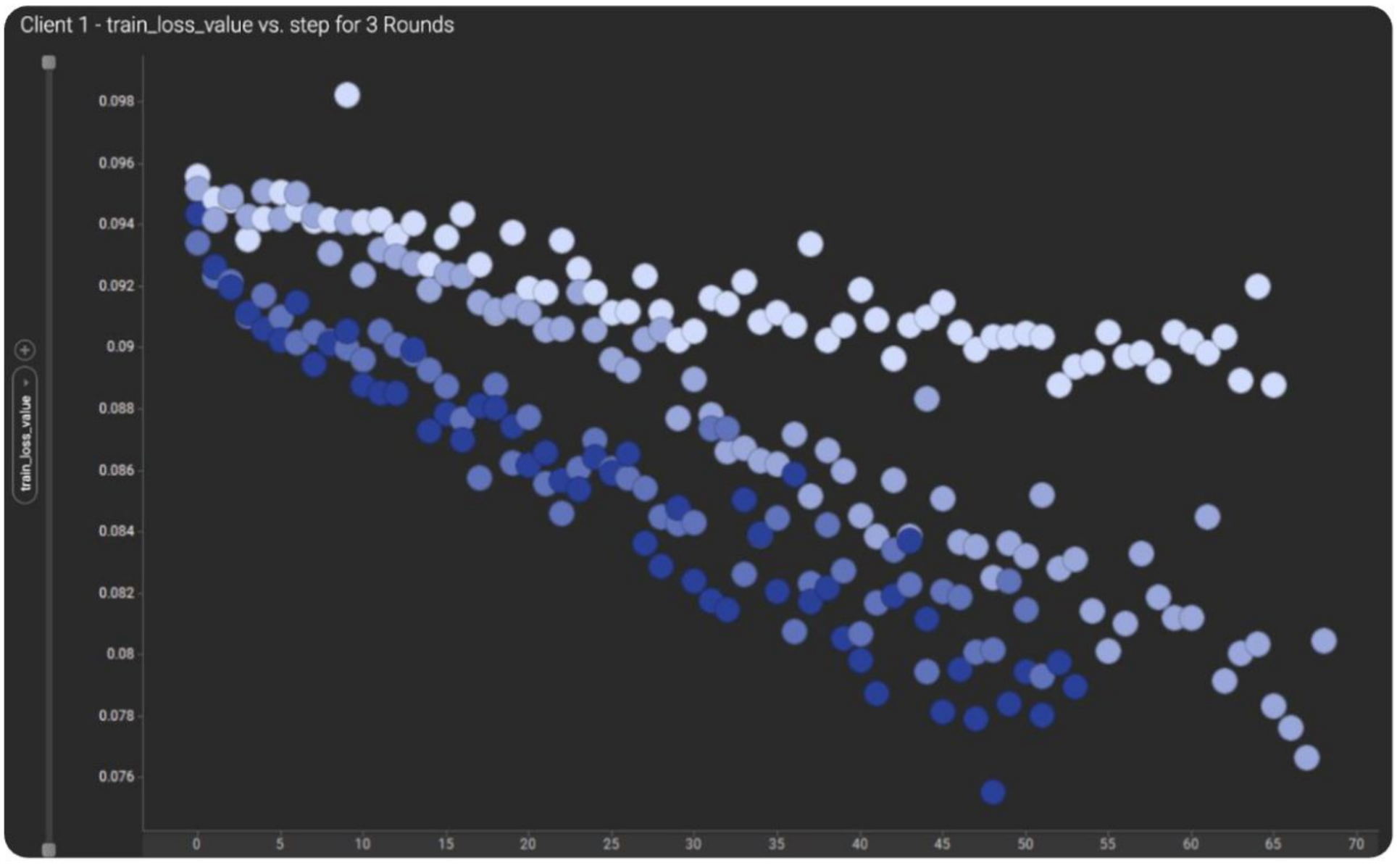


Figure 5: Client 1 training loss vs. step for three communication rounds. Bubble color represents communication round.

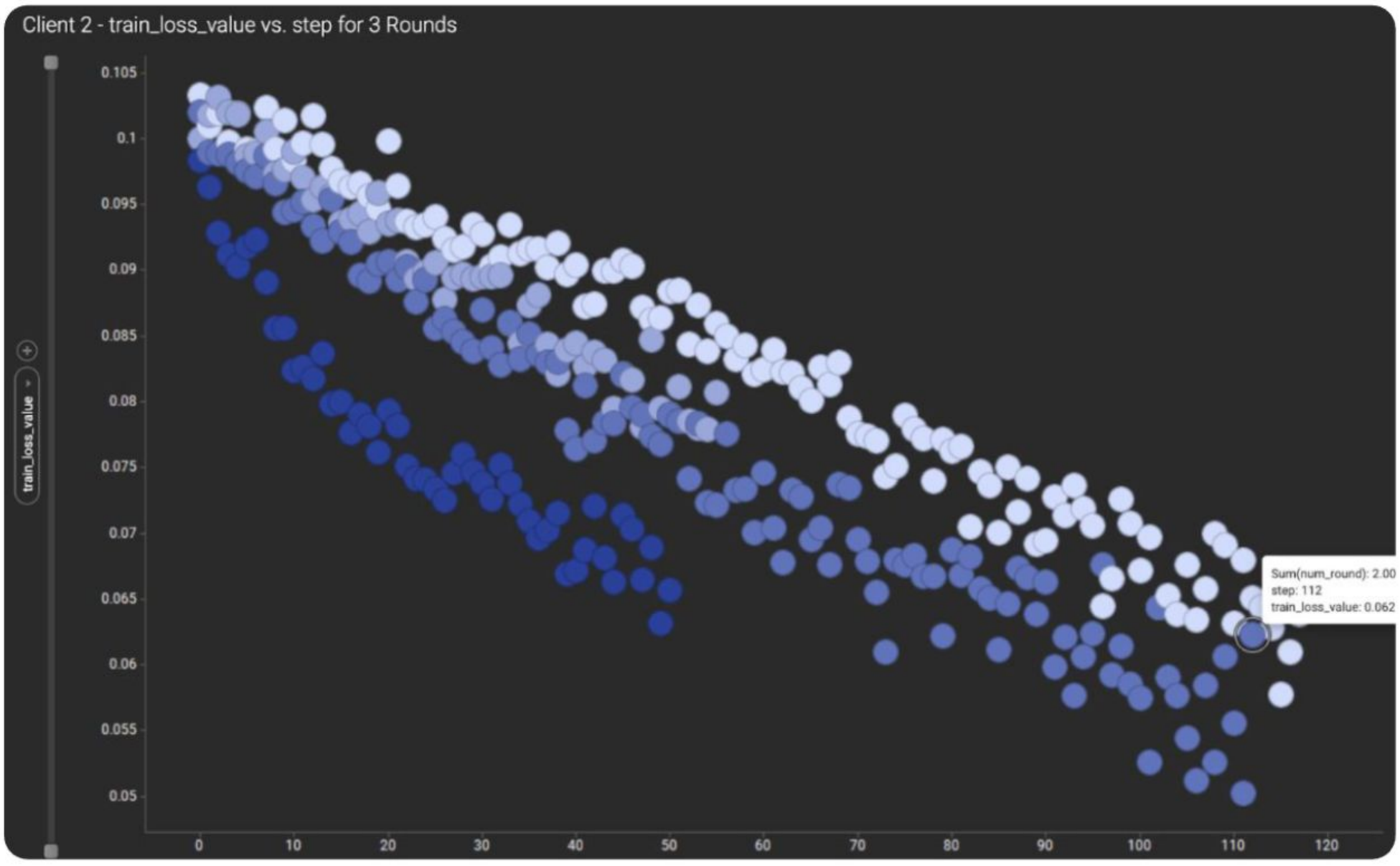


Figure 6: Client 2 training loss vs. step for three communication rounds.

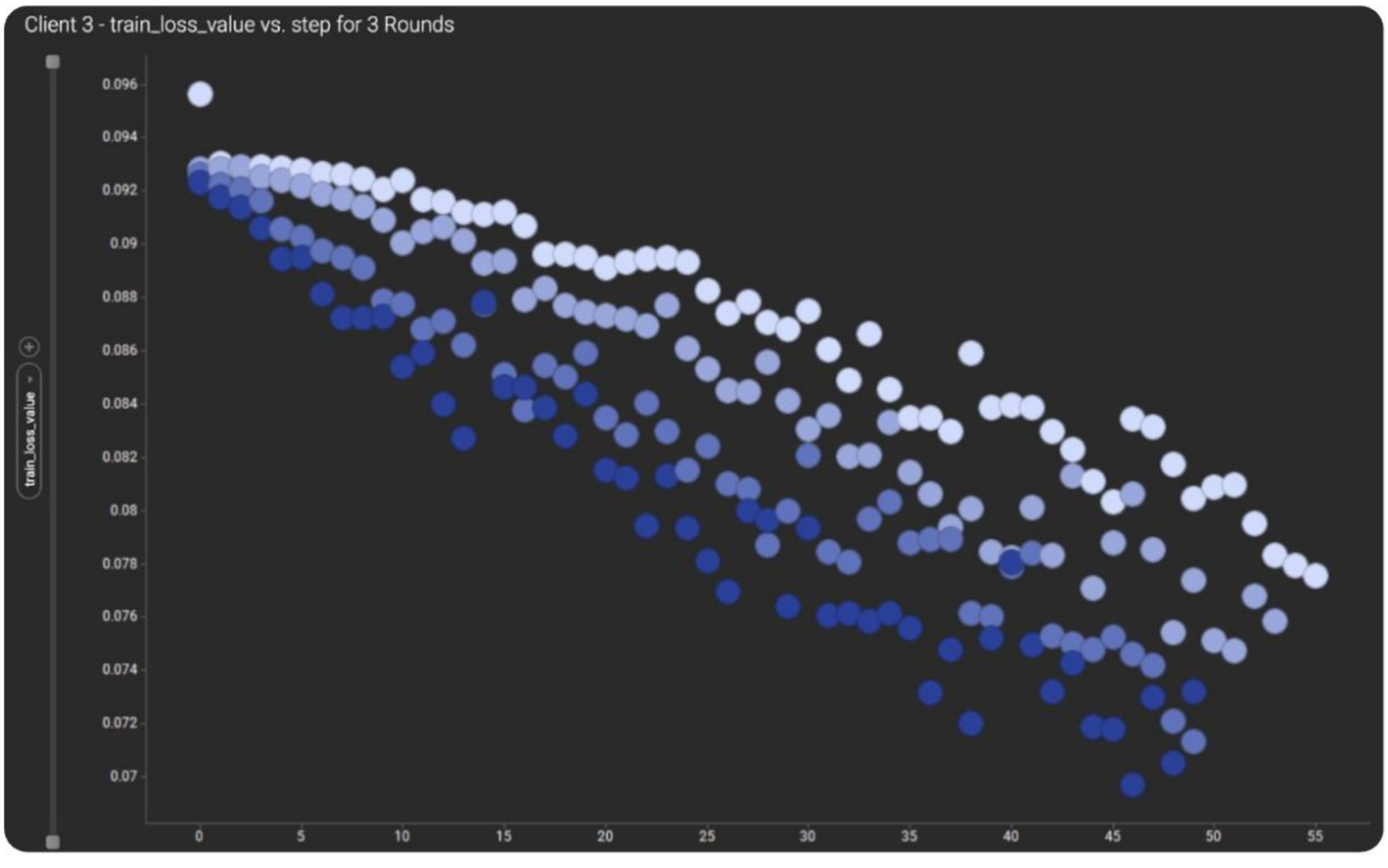


Figure 7: Client 3 training loss vs. step for three communication rounds.

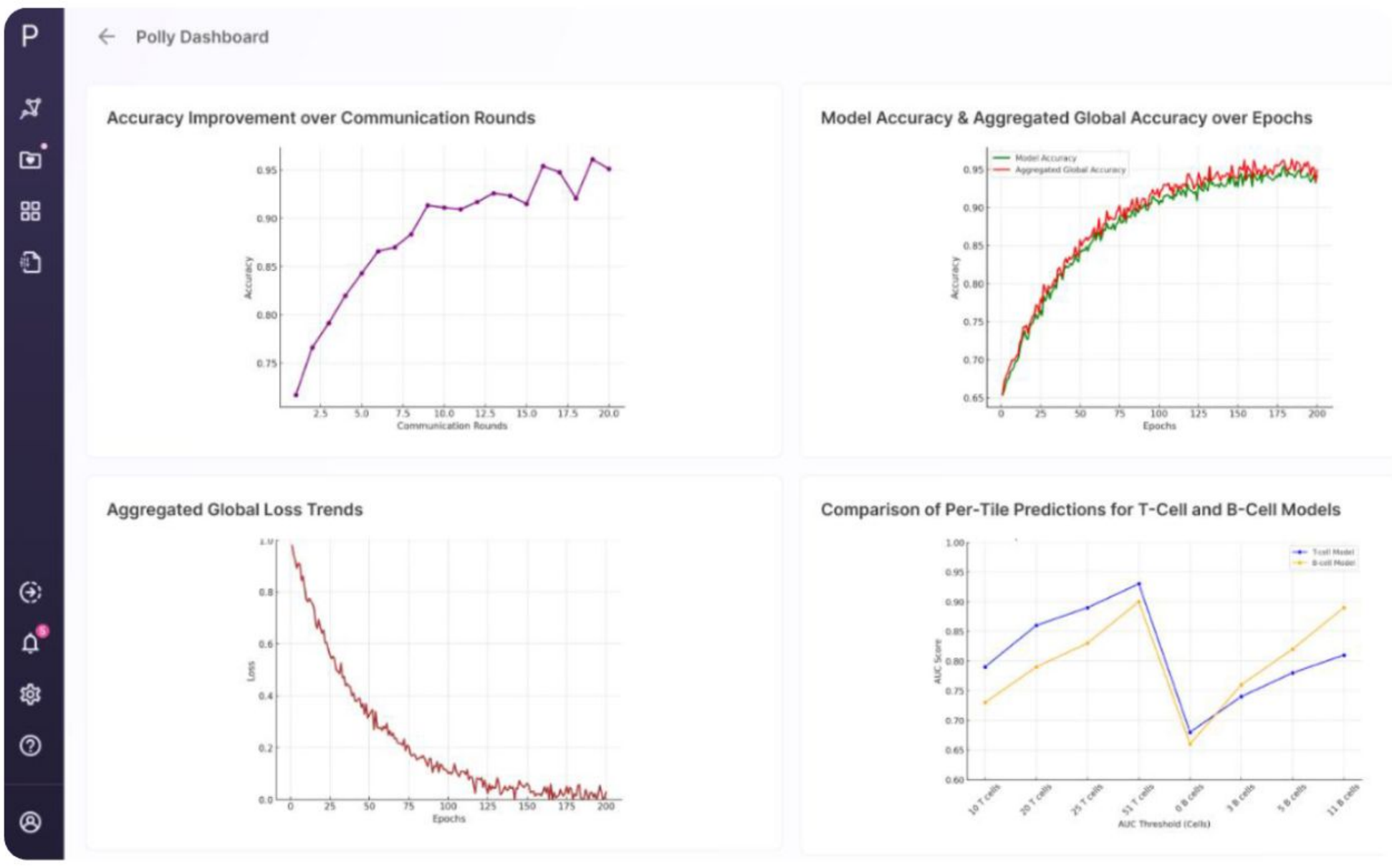


Figure 8: A page of Polly observability dashboard

Data Preparation and Curation

Data preparation represents the critical differentiator in biomedical federated learning success. Our domain specific preprocessing ensures model robustness across heterogenous data sources.

Whole Slide Image Processing Pipeline

Our bioinformatics team developed specialized WSI preprocessing for the gene expression prediction use case:

01 **Tissue Segmentation:** Custom U-Net models identify regions of interest, eliminate background artifacts, and direct computational resources on diagnostically relevant areas.

02 **Stain Normalization:** Macenko method implementation corrects inter-institutional variations in H&E staining protocols.

03 **Patch Extraction:** Sliding window approach generates 512×512-pixel tiles at 20× magnification, with intelligent sampling to balance tissue types.

Quality Assurance Framework

Maintaining data integrity across federated environments requires rigorous validation protocols. Our multi-tiered quality control system combines automated validation with expert oversight to ensure clinical-grade data reliability throughout the preprocessing pipeline.

- **Expert Review:** Our domain experts validate a 5% random sample of processed tiles, ensuring clinical relevance.
- **Metadata Integrity:** Checksums and provenance tracking maintain data lineage across distributed sites.

Multi-Modal Harmonization Capabilities

Biomedical research now demands integration across diverse data modalities, each of which poses unique preprocessing challenges. Our harmonization framework addresses the technical complexities of cross-modal data fusion while preserving the biological signal integrity essential for accurate federated learning outcomes. Beyond imaging, our platform supports:

- **Transcriptomics:** Batch effect correction using ComBat-seq for RNA-seq data.

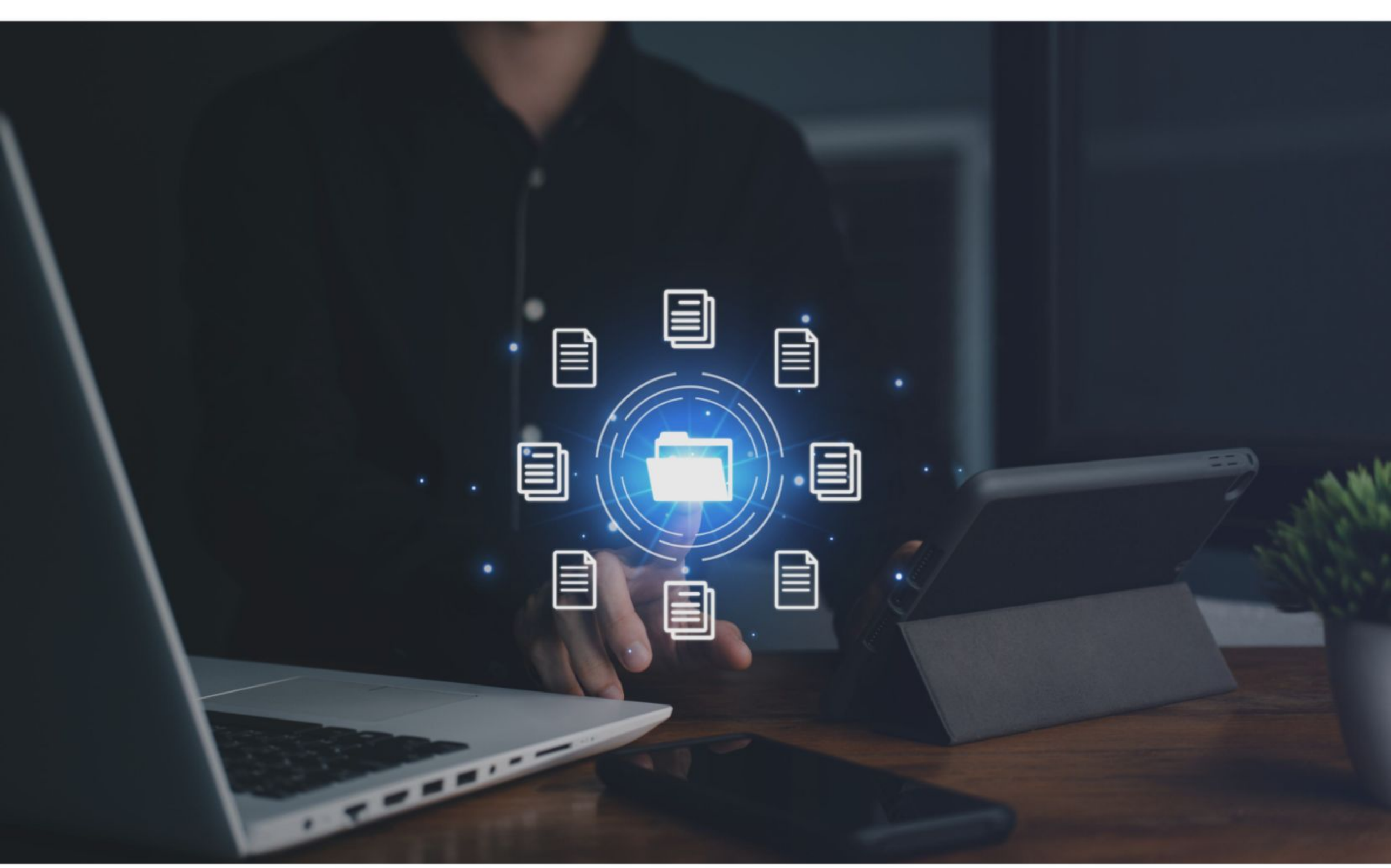
- **Proteomics:** Retention time alignment and intensity normalization for mass spectrometry.
- **Clinical Data:** OMOP CDM transformation for standardized phenotype representation.

Data Source Integration

Seamless connectivity across heterogeneous data ecosystems enables comprehensive biomedical analysis at scale. Our adaptive integration architecture works smoothly with both legacy systems and emerging data standards. It therefore ensures broad institutional compatibility while meeting all security and compliance requirements.

Our preprocessing pipelines integrate directly with:

- **Public Repositories:** TCGA, GEO, and ArrayExpress with automated metadata extraction.
- **Proprietary Datasets:** Custom adapters for institutional data warehouses.
- **Real-time Streams:** FHIR-compliant interfaces for EHR integration.



Strategic & Business Impact

Advancing Biomedical R&D Through Collaborative AI

Strategic Value of Elucidata's Federated Learning Implementation

Elucidata's Federated Learning platform transforms biomedical research by unlocking new possibilities for both accelerated discovery and strategic collaboration. By removing traditional data sharing barriers, organizations can now access unprecedented opportunities for innovation while maintaining full control over their proprietary assets.

Accelerating Discovery for R&D Teams

Our solution fundamentally changes the economics of biomedical AI development by enabling organizations to leverage distributed data resources without the traditional barriers of data centralization. This paradigm shift creates unprecedented opportunities for accelerated research and development across the industry.

Key advantages of Elucidata's FL solution include:

- **Model Diversity:** Access to 10-100x more diverse training data without data acquisition costs.
- **Reduced Bias:** Models trained on global populations improve efficacy predictions across ethnic groups.
- **Faster Iteration:** Parallel training across sites reduces time-to-insight by 60-75%

Novel Collaboration Models

Previously impossible partnerships are enabled through this FL solution, by removing the fundamental barrier of data sharing while maintaining institutional control and privacy. This creates new opportunities for collaborative research that were previously hindered by competitive concerns and regulatory constraints.

Emerging collaboration opportunities:

- **Competitive Collaboration:** Pharmaceutical companies jointly train models while protecting proprietary compound libraries.
- **Academic-Industry Bridges:** Universities contribute data without intellectual property concerns.
- **Global Health Initiatives:** Low-resource settings participate without expensive infrastructure requirements.

Elucidata's Differentiated Solution Benefits

It also delivers transformative business value that extends far beyond technical capabilities, and creates sustainable competitive advantages across regulatory, operational, and strategic dimensions. Our comprehensive approach not only solves today's data collaboration challenges but positions organizations as pioneers in the responsible AI ecosystem of tomorrow.

Regulatory Compliance as Competitive Advantage

Our privacy-preserving architecture transforms compliance from a barrier to an enabler, positioning Elucidata clients at the forefront of regulatory innovation. By building privacy protection directly into our federated learning infrastructure, we enable organizations to exceed compliance requirements while accelerating their research timelines.

The comprehensive compliance coverage of Elucidata's solution includes:

- **GDPR:** Data protection by design satisfies stringent EU requirements
- **HIPAA:** De-identification occurs at source, eliminating PHI transmission risks
- **SOC 2 Type II:** Continuous monitoring ensures ongoing compliance certification

Emerging collaboration opportunities:

- **Competitive Collaboration:** Pharmaceutical companies jointly train models while protecting proprietary compound libraries.
- **Academic-Industry Bridges:** Universities contribute data without intellectual property concerns.
- **Global Health Initiatives:** Low-resource settings participate without expensive infrastructure requirements.

Risk Mitigation and Time-to-Market Acceleration

Both technical and business risks associated with large-scale biomedical AI initiatives are significantly reduced by utilizing Elucidata's FL Platform. Our approach enables organizations to validate concepts and build models without the substantial upfront investments traditionally required for centralized data infrastructure.

Strategic advantages include:

- **Infrastructure Reuse:** Eliminate 6-12 months of data center setup by leveraging existing institutional resources.
- **Reduced Legal Overhead:** Standard FL agreements replace complex multi-party data use agreements.
- **Fail-Fast Capability:** Validate model feasibility before committing to full data centralization.
- **Regulatory Pre-Approval:** Demonstrate privacy preservation to regulators before study initiation.

Quantifiable Competitive Advantages

Elucidata's federated learning implementation Once implemented, this solution has quantifiable competitive advantages. It delivers measurable business value that directly impacts our clients' bottom line and competitive positioning. These quantifiable benefits demonstrate clear ROI while positioning organizations as leaders in responsible AI development.

Measurable outcomes:

- **Competitive Collaboration:** Pharmaceutical companies jointly train models while protecting proprietary compound libraries.
- **Academic-Industry Bridges:** Universities contribute data without intellectual property concerns.
- **Global Health Initiatives:** Low-resource settings participate without expensive infrastructure requirements.

Future Enhancements

Evolution Towards Autonomous Federated Learning

Our federated learning platform is designed for enterprise-scale deployment with robust performance characteristics. Scalability is achieved through minimal architectural changes required to add new participants, which supports collaborative efforts at regional, national, or even global scales. Performance is optimized via S3 Gateway Endpoints and private AWS infrastructure, ensuring low-latency, high-throughput data access for distributed training across geographically dispersed institutions.

While our current work is production-ready, we are continuously evolving our solution to keep it on the edge of technological advancements and improve ease of use.

This roadmap encompasses both immediate operational improvements and breakthrough research directions.

Operational Enhancements

The platform's future operational development focuses on streamlining deployment processes and extending accessibility across diverse computing environments. The following enhancements aim to reduce technical barriers for research institutions while enabling seamless scaling from high-performance computing clusters to resource-constrained edge devices in clinical settings.

- **Plug-and-Play Deployment** via operators for single-command FL cluster provisioning.
- **AutoML Integration** for hyperparameter optimization across federated networks.
- **Cross-Silo to Cross-Device** extension from institutions to edge devices (wearables, point-of-care diagnostics).

Research and Capability Expansion

Scientific advancement requires continuous expansion of the platform's analytical capabilities to accommodate emerging data types and modeling paradigms in biomedical research. The research roadmap prioritizes the integration of cutting-edge genomics technologies and the development of specialized foundation models tailored for healthcare applications. It includes:

- **Expanded Modality Support,** including spatial transcriptomics and single-cell multiomics preprocessing
- **Federated Foundation Models** for pre-training large vision-language models in biomedical applications

Security Roadmap

Homomorphic encryption keeps model parameters encrypted throughout processing, allowing calculations without decryption. This is suitable for ultra-sensitive biomedical projects where even aggregated updates must remain protected.

Adjacent Problem Spaces

Apart from use cases discussed above, Elucidata's solution for federated learning will also help in other adjacent spaces, some of which are:

- Elucidata's FL solutions can be applied to **Clinical Trial Optimizations,** e.g., Federated site selection and patient stratification.
- This FL solution can be deployed **for Real-World Evidence Generation** using continuous learning from distributed EHR systems.
- **Precision Medicine Networks,** e.g,. Federated pharmacogenomics for personalized dosing.
- **Global Pandemic Preparedness** by rapid model development across international borders.

Conclusion

Biomedical AI has reached an inflection point. Stricter privacy rules, exploding multimodal data, and urgent clinical needs now outstrip centralized learning. The chief obstacle is no longer algorithms but the impracticality of pooling sensitive biomedical data at scale.

Elucidata's federated learning platform, enhanced by deep biological domain expertise and the Polly observability layer, offers a proven path forward.

Our case study demonstrates that institutions can achieve superior model performance while maintaining complete data sovereignty. We transform regulatory compliance from an obstacle into an accelerator for innovation by keeping data local and sharing only learned insights.

Elucidata's disruptive approach addresses three critical market dynamics:

Regulatory Momentum

New frameworks explicitly encourage privacy-preserving technologies.

Technical Maturity

Production-ready infrastructure eliminates previous adoption barriers.

Competitive Pressure

Early adopters will capture disproportionate value from collaborative AI.

Organizations **adopting Elucidata's federated learning platform today will achieve superior AI model accuracy** through diverse, collaborative training while maintaining complete data privacy, delivering faster regulatory approvals, and a decisive competitive edge in the $600+ billion healthcare AI market.

Early adopters won't just participate in the precision medicine revolution; they'll lead it by setting the new standard for trustworthy, collaborative healthcare AI.

Your Next Steps

- **Schedule a Technical Deep-Dive with us:** Explore how federated learning applies to your specific use cases
- **Access Our Technical Resources:** Download our whitepapers, go through our technical blogs

Contact our team at harshveer.singh@elucidata.io to begin your journey toward privacy-preserving collaborative AI.

References

[1] IBM Security. (2023). Cost of a data breach 2023: The healthcare industry. IBM Corporation. https://www.ibm.com/reports/data-breach

[2] IDC & Seagate Technology. (2018). The digitization of the world – From edge to core [White paper]. IDC. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[3] Precedence Research. (2024). Artificial intelligence in healthcare market size to hit USD 613.81 bn by 2034. https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market

[4] MedCity News. (2023, October 10). Health Exec: 97% of Healthcare Data Isn't Used. Retrieved from https://medcitynews.com/2023/10/health-exec-97-of-healthcare-data-isnt-used/

[5] Li, T., Sahu, A. K., Talwalkar, A. and Smith, V. (2020) 'Federated learning: challenges, methods, and future directions', IEEE Signal Processing Magazine, 37(3), pp. 50–60. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7614889/ (Accessed: 9 June 2025).

[6] Press, G. (2016) 'Data preparation most time-consuming, least enjoyable data science task, survey says', Forbes, 23 March. Available at: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (Accessed: 9 June 2025).

[7] Schmauch, B., Romagnoni, A., Pronier, E. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun* 11, 3877 (2020). https://doi.org/10.1038/s41467-020-17678-4

[8] IJCAI PDF n.a. (2022). Replace with exact title/first author of the paper. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22). International Joint Conference on Artificial intelligence https://www.ijcai.org/proceedings/2022/0791.pdf

[9] AWS blog post Amazon Web Services. (2020, May 13). Machine learning with decentralized training data using federated learning on Amazon SageMaker. AWS Machine Learning Blog. https://aws.amazon.com/blogs/machine-learning/machine-learning-with-decentralized-training-data-using-federated-learning-on-amazon-sagemaker/

[10] Flower website Flower. (n.d.). Flower – a friendly federated-learning framework. Retrieved 30 May 2025, from https://flower.ai/

[10] Schmauch, B., Romagnoni, A., Pronier, E. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun 11, 3877 (2020). https://doi.org/10.1038/s41467-020-17678-4

About Elucidata

Elucidata is a San Francisco and Boston-based deep tech company which enables organizations to become AI-ready. Powered by a 110+ strong multidisciplinary team, the company's technology delivers AI-ready data 10X faster with 60% reduction in data processing costs through automated workflows. Keeping data quality at its foundation, Elucidata's proprietary data quality checks ensure 99% accuracy in multi-modal data through a human-in-the-loop approach. Rooted around people, technology and process, Elucidata's mission is to keep AI 'Real' for precision medicine.

Have questions or want to discuss the ideas shared in this whitepaper?

We'd be happy to hear from you.

For any queries, clarifications, or to explore how these approaches could apply to your work, please reach out to Harshveer Singh at harshveer.singh@elucidata.io

Book a Demo