Breaking Barriers, Not Privacy: Real-World Split Learning Across Healthcare Systems

By

Sravan Kumar Elineni

Authors: Sean Muir, Dave Carlson, Jayme Welty, Keith Salzman,
Serafina Versaggi , Lindsay Cushing, Sravan Kumar Elineni – BookZurman Inc
Himali Saitwal, Savannah Mueller - EMI advisors
David E Taylor - Regenstrief
Chit win – IOI/MIHIN
Adam Wong, Jordon Everson – ASTP/ONC

Agenda

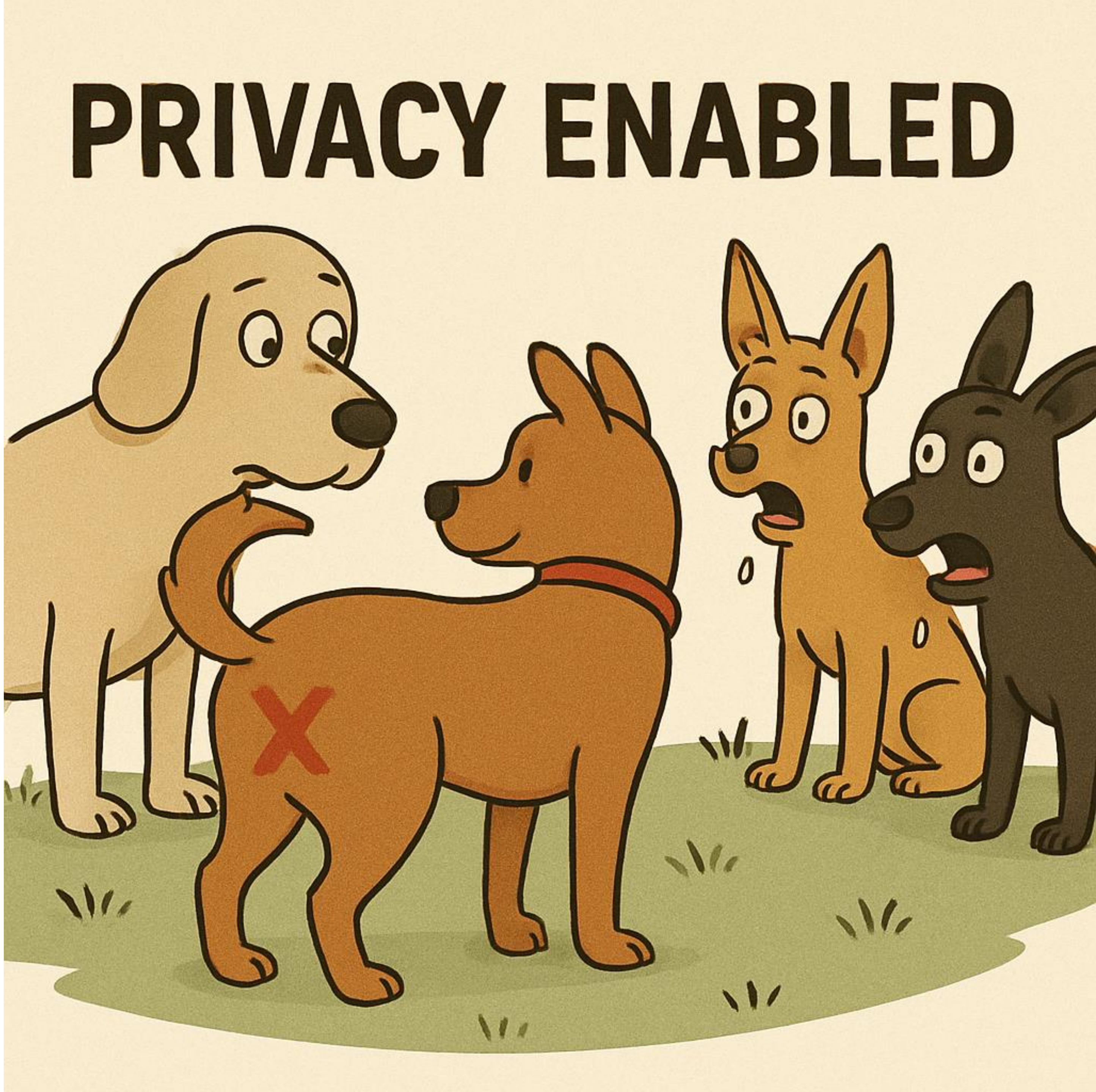
01 Project outline- What, Why, How

02 Data and interoperability

03 Technical and security overview

04 Real world challenges

05 Future direction



Outline

What and why of the project

Healthcare Data Challenge

- Decentralized data across institutions limits collaborative research

Privacy Barriers

- Regulations restrict data movement between organizations

Collaborative AI with Privacy

Standards, data governance and privacy preserving deep learning are key for collaboration

Standards & interoperability → **CONSISTANCY**

- ASTP's 21st Century Cures Act
- USCDI FHIR and bulk FHIR api
- Interoperability in demographics, labs, medications etc.

Data governance → **STANDARDIZE**

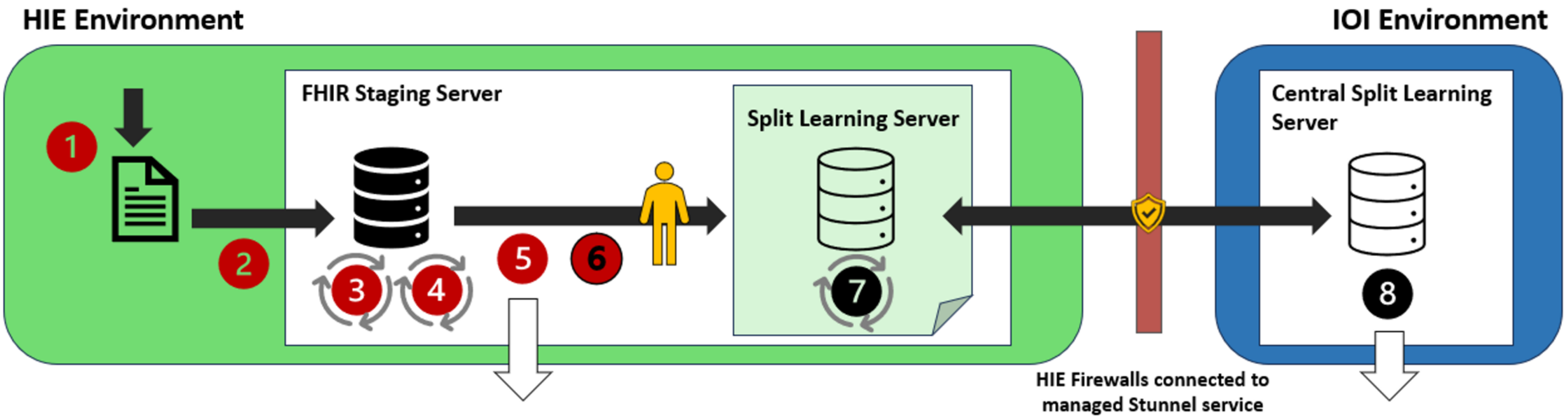
- NIH's VSAC value sets , MCC authored value sets
- Project specific value sets(40 conditions, 29 labs etc.)
- CQL for querying FHIR enforcing value sets
- Consistent input into data loader from each HIE

Secure framework → **PRIVACY**

- Split learning
- Federated learning etc.

Project Process Flow

Multi modal data is processed, aligned into encounters using FHIR-CQL-valuesets before ingesting to SL



FHIR Bundle	Honest Data Broker Review	Split Learning File(s)	(Black) Data has PHI Removed
FHIR Server	(Red) Data contains PHI	Data Flow	Outputs

Multi Modal Data

Multi modal data is processed, aligned into encounters using FHIR-CQL-valuesets

Continuous variables

Categorical variables

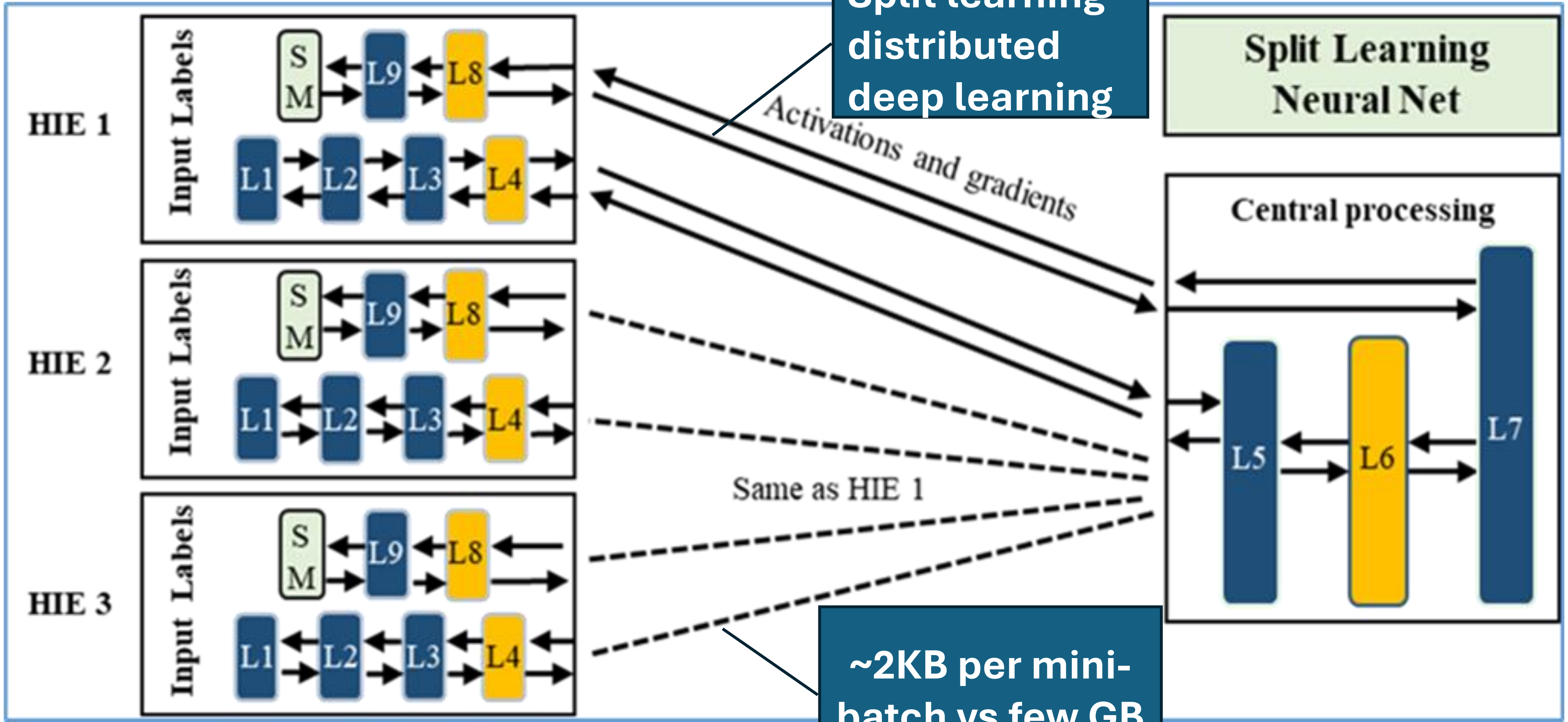
Notes

- 40 conditions, 9 labs and 5 symptoms from MCC value sets
- 20 additional lab value sets like arterial blood gas etc.

- Deidentified and obfuscated
- embeddings of size 768

Normalized, standardized input to neural networks

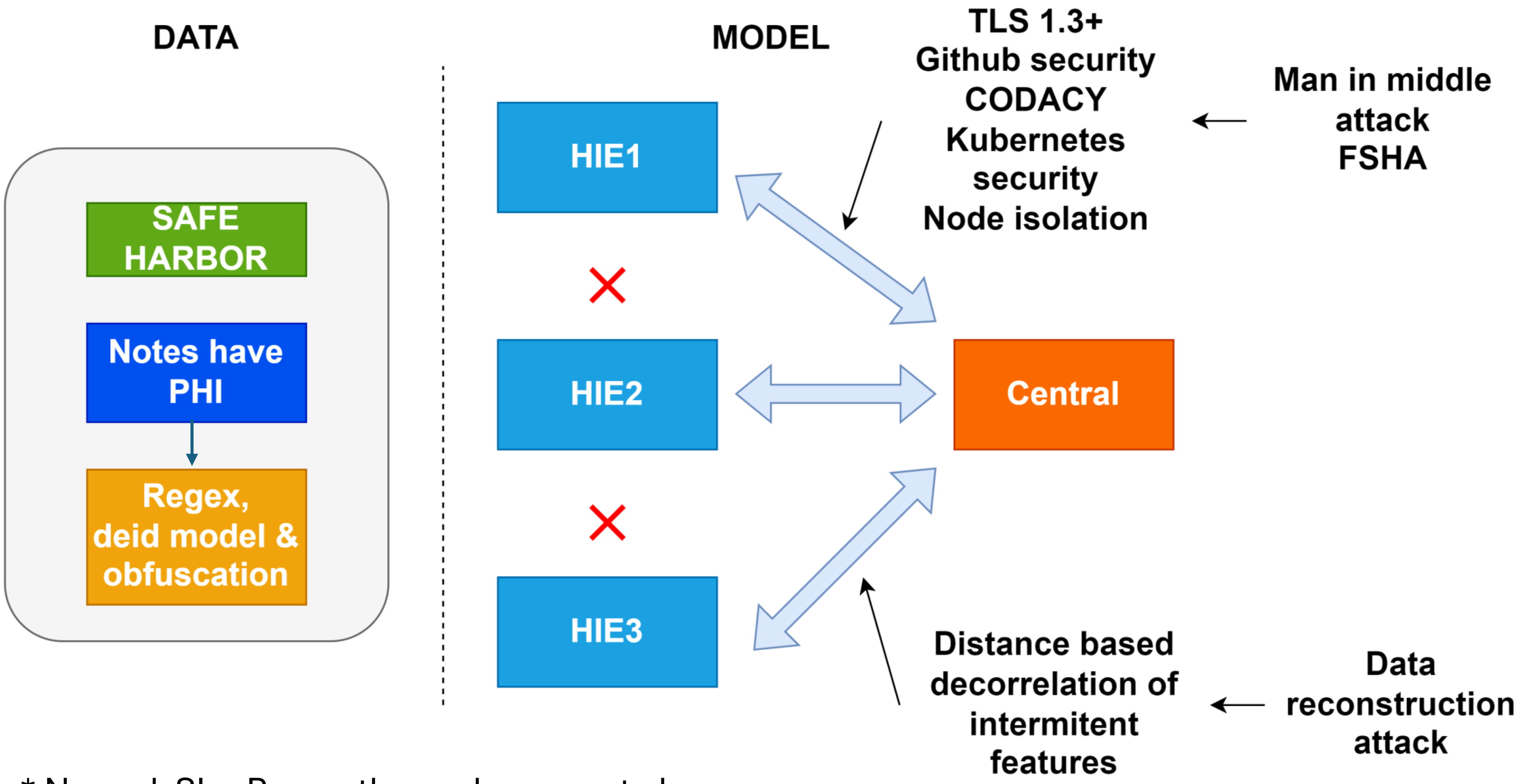
Split Learning Architecture Overview



Split learning distributed deep learning

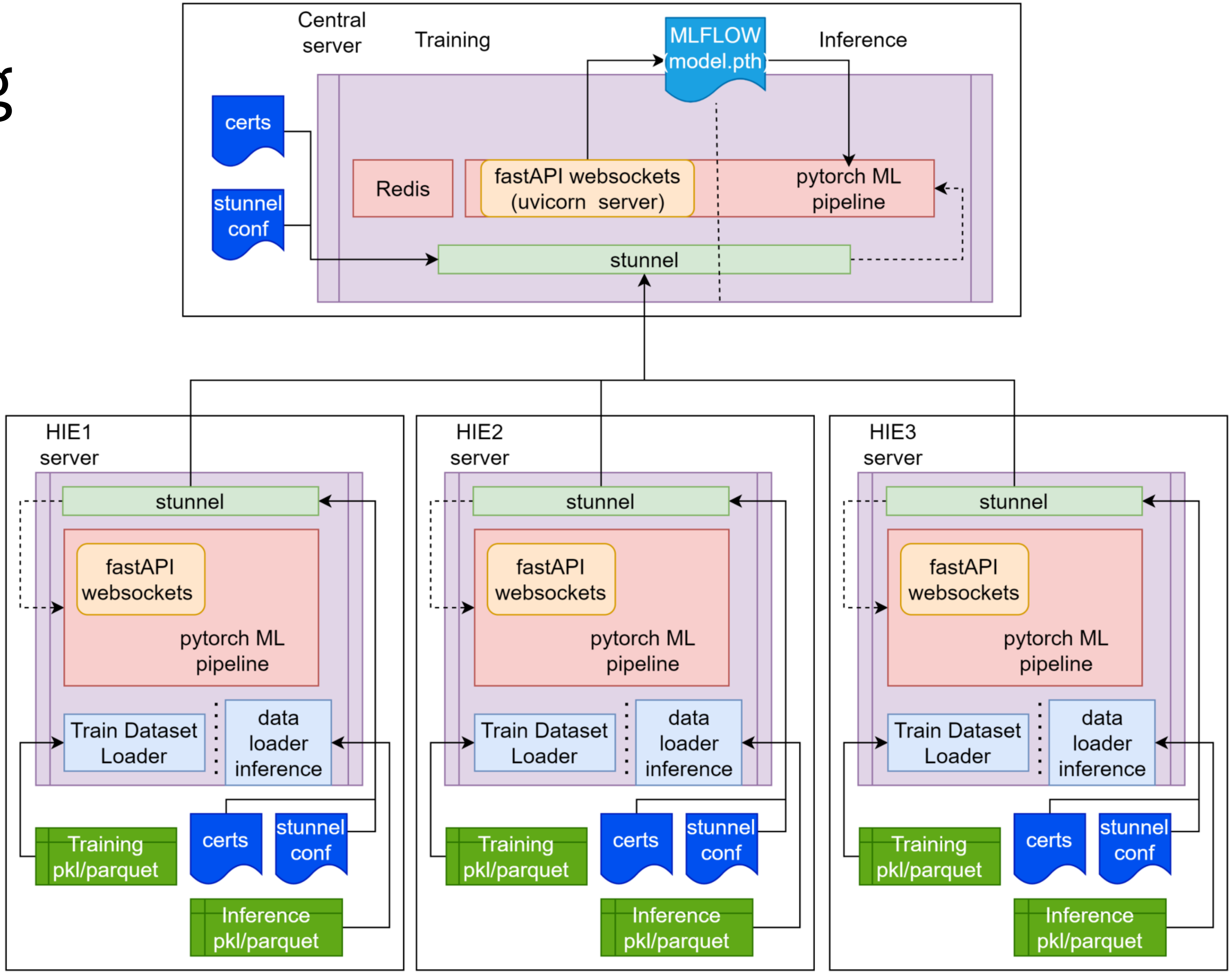
~2KB per mini-batch vs few GB

Privacy Engineering Decisions (semi-honest)



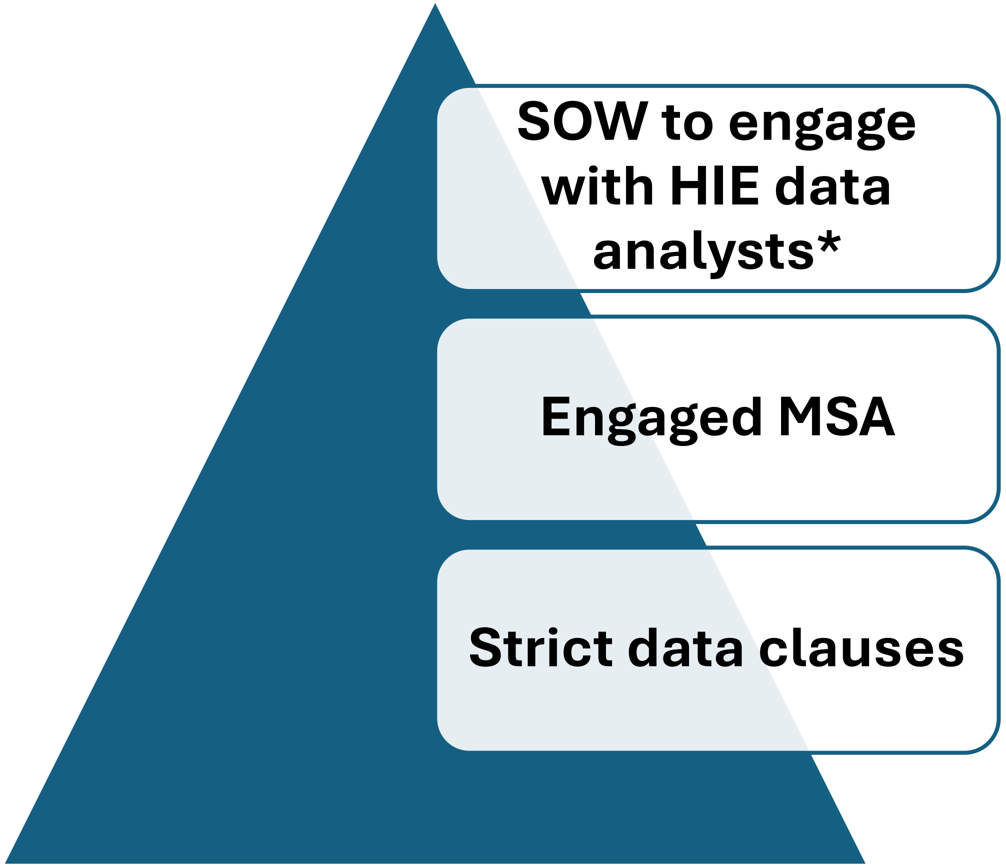
* Nopeek SL – Praneeth vepakomma et al.

Split Learning Architecture



HIE Collaboration Challenges

Data clauses clearly defined roles and restrictions for each party



*Strict data use clause meant the central server research team had no visibility into patient data. Successful EDA and mapping correction required healthcare domain expertise and the active role of HIE-side analysts.

Data Quality Verification

Data quality is verified at aggregate level

FHIR Data Mapping Strategies

- Automated quality summaries without PHI exposure
- Code coverage and terminology alignment analysis
- Real-time FHIR resource utilization feedback

Handling Inconsistent Implementations

- Standardized LOINC codes for lab data
- Filtered non-affirmative diagnostic codes
- Resolved unit measurement discrepancies

QA Check Methodology vs CDC Data

- Validated comorbidity rankings against CDC Indiana data
- Confirmed heart disease as top COVID-19 comorbidity
- Verified alignment across diabetes, kidney, cognitive disorders

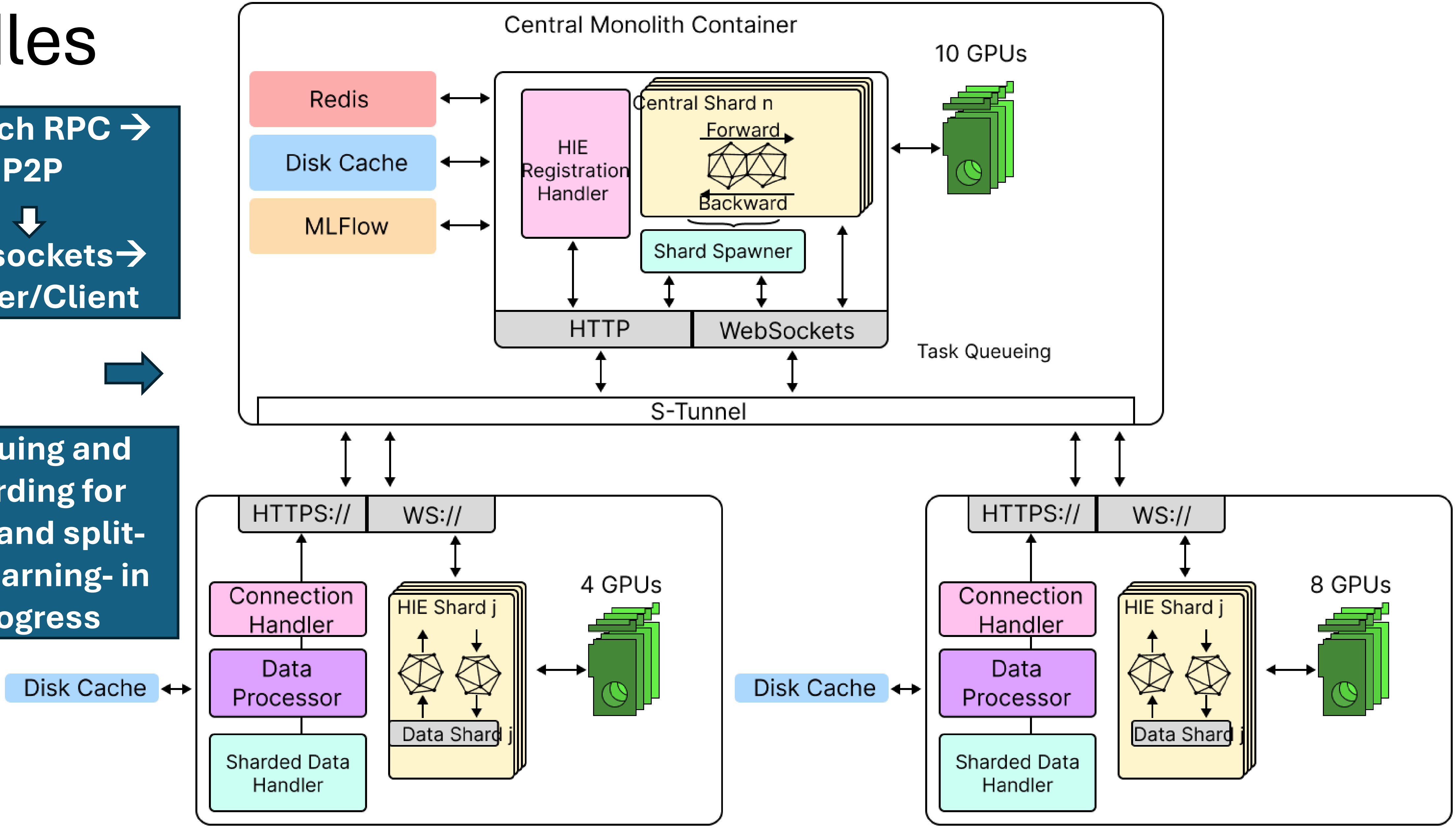
Technical Implementation Hurdles

1 **Pytorch RPC → P2P**

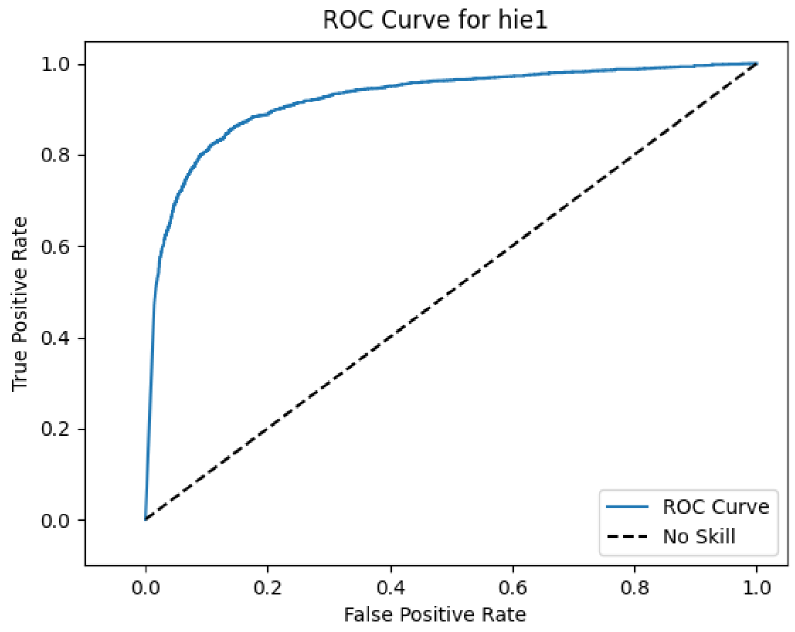
⇓

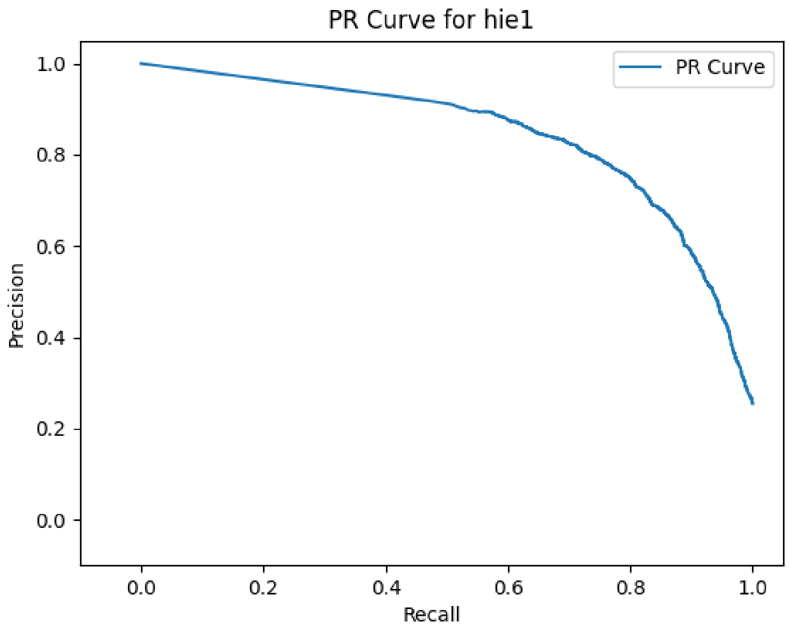
Websockets→ Server/Client

2 **Queuing and sharding for split and split-fed learning- in progress**



Results and Dissemination





- **Worked on a multi modal model for disease diagnosis (5 comorbidities)**
- **Utilized categorical, continuous and clinical text features**
- **With 32:16 batch size for train:test, we have Precision@0.8 and recall@0.74**

Future Direction

Multi-GPU: Optimization for split learning workflows to accelerate privacy-preserving model training across distributed HIE systems on multiple GPU

Advanced Security: Differential privacy-inspired activation function at splits, real-time privacy scoring, and attack-vector testing

Model Enhancements: Positional encodings for temporal data, FocalBCELoss for class imbalance etc

Let us Collaborate to refine split learning across healthcare research for broader impact

Why split learning?

Primary

- Speed. About 2kb is usually passed across internet vs GB of model data

Secondary

- Fine grained control over modelling to control privacy, learning process etc

CDC report comparison to HIE data

Adverse Conditions RI HIE	Ranking	Adverse conditions from Indiana
Heart Disease	1	Heart Disease
Diabetes	2	Cancer
Kidney Disease	3	Accidents
Cognitive Disorder	4	Chronic Lower Respiratory Diseases
Stroke	5	Stroke
	6	Diabetes
	7	Alzheimer’s Disease
	8	Kidney Disease
	9	Chronic Liver Disease/Cirrhosis