Secure AI Labs White Paper

Novartis PoC with Federated Learning to Produce Microbiome Findings

Maurice Lubetzki, Matthieu Pichaud, and Anne Kim

July 2019

Executive Summary

SAIL worked with a principal investigator at Novartis to replicate previous genomic findings using SAIL's Federated Learning technology. The replicated results demonstrated a level of accuracy comparable to the original results. Furthermore, using Federated Learning provides more security benefits.

- Principal Component Analysis results showed high similarity and no technical or biological bias;
- Wilcoxon Signed-Rank Test showed similar results in the Federated environment compared with the original results.

Introduction

The purpose of this study was to test Secure AI Labs (SAIL) privacy technology on microbiome data held by hospitals. Open data sharing is one of the biggest bottlenecks to research in healthcare. The SAIL solution is designed to keep data in its original server while being analyzed in a cryptographically private environment by external analysis. To execute this, SAIL leverages enclave technology and federated learning.

Definition of Federated Learning

"A machine learning setting where the goal is to train a high-quality centralized model with training data distributed over a [...] number of [privacy sensitive data providers]."[1]

Definition of Secure Enclave

A secure enclave is "an isolated memory location" that "[protects] applications and data at runtime", meaning it runs in a "trusted execution environment (TEE)."[2]

The study was led by a principal investigator at Novartis in Boston who provided a genomic data set based on stool samples coming from four separate hospitals. Each patient in the data set is associated with genomic data and a diagnosis that's ulcerative colitis

[1] https://ai.google/research/pubs/pub45648

[2] https://www.infosecurity-magazine.com/opinions/enclaves-security-world/

(UC), Crohn's disease (CD), and non-inflammatory bowel disease (IBD). We normalized analysis across the four hospitals and verified that the distributions are similar for comparison by running a principle component analysis (PCA). Once we confirmed comparability, we ran a Wilcoxon signed-rank test to determine which microbes are most significant for UC, CD, and IBD.

Methods & Results

Principal Component Analysis (PCA)

Across four hospitals, we computed a differentially private[3] federated learning PCA on the genomic dataset. After appropriate data normalization, variance covariance matrices were computed across each data set within four secure enclaves, one for each hospital. Prior to sending the matrices, we summed them with noise to ensure differential privacy.

Figure 1: Traditional (vulnerable) data transfer

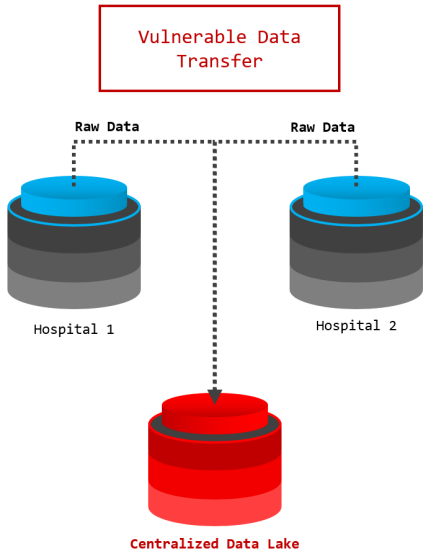
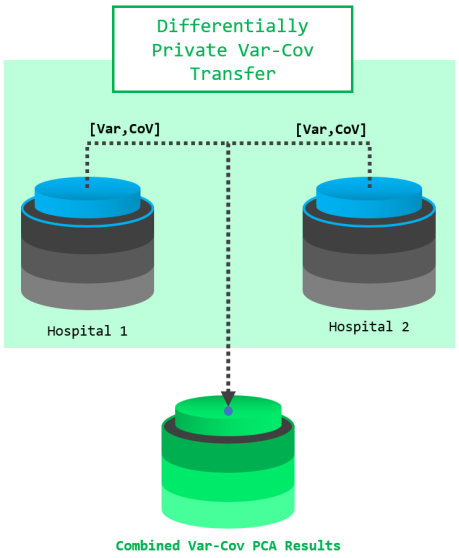


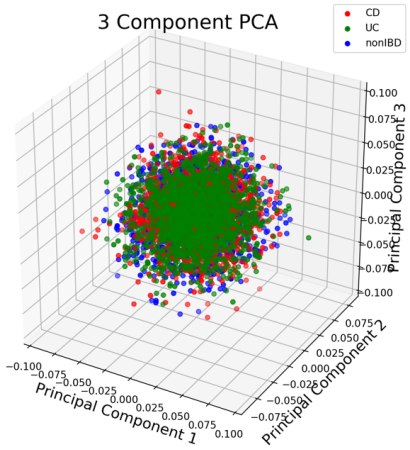
Figure 2: Differentially private (secure) data transfer



This noise is only a matrix sampled from a multivariate normal distribution with its standard deviation scaled relatively to the size of the data (Fig. 1 and 2). The hospitals' matrices were gathered in a central enclave where they were summed up prior to the PCA being executed on the latter matrix. Finally, principal components are gathered for each diagnosis.

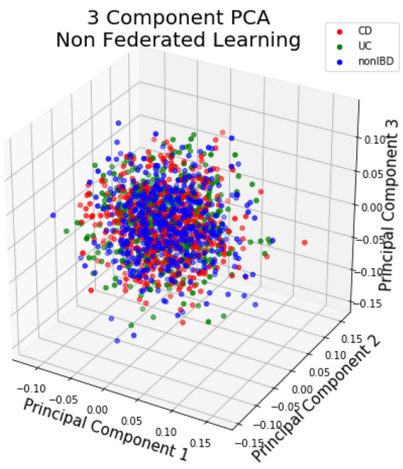
Figure 2: Federated vs. non-federated PCA results demonstrate that federated learning can produce accurate models

Figure 2A: Federated PCA



[3] https://en.wikipedia.org/wiki/Differential_privacy

Figure 2B: Non-federated PCA



After plotting the three principal components federated and non-federated wise, we noted a high similarity in the diagnosis principal components distribution regardless of the machine learning setting implying a strong conclusion that there is neither technical nor biological bias. After getting confirmation from the PCA results that there was no bias between the four hospital datasets, we concluded the datasets were comparable for further analysis. Thus, we proceeded to the Wilxocon Signed-Rank Test.

Wilcoxon Signed-Rank Test

Researchers at the pharmaceutical company wanted to understand which microbes were most important to which specific disease. We implemented a Wilcoxon Signed-Rank Test in normal and SAIL environments. We found that ranking results done in our cryptographic environment were comparable. Furthermore, the speed penalties were also within reasonable tolerance (10 minutes instead of 3 minutes).

MSP Name	rank1		rank2		Statistic		P Value	
	Normal	SAIL	Normal	SAIL	Normal	SAIL	Normal	SAIL
msp_0001	8	8	13	12	-0.96	-0.72	0.34	0.47
msp_0002	2	2	8	8	-1.20	-1.19	0.23	0.24
msp_0003	28	28	32	32	0.51	0.67	0.61	0.51
msp_0004	3	3	6	5	-0.57	-0.36	0.57	0.72
msp_0005	28	28	33	33	-0.74	-0.61	0.46	0.54
msp_0006	1	1	2	2	-0.15	-0.15	0.88	0.88
msp_0007	28	28	30	30	1.24	1.35	0.21	0.18

Table 1: SAIL demonstrates ability to replicate accurate results using Federated Learning methodology

Extensions & Implications

Ultimately, the conclusion of this proof-of-concept is that the SAIL platform is well suited for distributed data assessment (PCA) and analysis (rank sum for microbes of interest). The conclusions extend beyond microbiome analysis into behavioral, multi-omics, chemical, eQTL, and other clinical interests in the pharmaceutical space where timely access to the most of the right data is essential to furthering research. Data accessibility in silo searching, internal review boards (IRB), anonymization, and transportation can take anywhere between 20-30% of a project's timeline because of restrictions in privacy, security, and compliance. The vision of the SAIL platform is to streamline this process with a platform that makes privacy, security, and compliance seamless at the software level, finally fulfilling the gap between written contract agreements to the very execution of code on the platform.

Secure AI Labs Solution

The Problem

Research is road blocked by timely access to the right data. The problem unfolds in the following problematic steps:

1. Data is siloed across different companies and across departments in the same company [1].
2. In the current research paradigm of extricating data, internal review boards (IRB) will take months—or even years to assess the methods and risk of a study.
3. The guidelines of IRB's often enforce data anonymization that inherently removes detail and often sacrifices analysis precision. Furthermore, anonymization doesn't even completely protect the data from being identified [2, 3].
4. Because the data is so large, the fastest file transfer solution thus far is to literally mail hard drives—hardly a secure solution. To avoid the liability, data should not be moved.
5. When data is behind a collaborator's firewall, it is difficult to audit all operations on company data. Currently there is a severe technical gap between what companies agree upon in IRB's and what actually happens behind the purview of a collaborator's firewall.
6. Once data has been moved, it is very tempting for IT teams to deposit all the shared data in a shared data lake, which is a major security vulnerability.
7. Alternatively, any purely federated approach in which algorithms are moved and deployed in order to ensure data security, does not protect the algorithm. Algorithms not only have proprietary methodology but also have proprietary data baked into the model parameters. In fact, with as few as 41 queries (or prediction tasks by the algorithm), anyone can steal the algorithm with 98% accuracy [4].

Solution

The Secure AI Labs solution uses four technologies to address these problems

1. Federated Learning – protects data (addressing problems 1,2,4,6). Federated learning is the machine learning method of sending algorithms to the dataset to learn locally and then aggregating the learnings for analysis on disparate datasets [6,7].
2. Differential Privacy – protects data (addressing problem 3). Differential privacy is the process of adding noise to the aggregated learnings to further protect data samples. Unlike other anonymization methods, differential privacy shifts the paradigm of privacy from a crude metric based on entire column definitions but instead by aggregate privacy based on mathematically defined noise unique to each dataset [7].
3. Secure Enclaves – protects algorithms (addressing problems 5,7). Enclaves provide vetted hardware that assures that a computational environment in which an algorithm is secured is safe and audited. This is based on pre-deployed protocols that are standard to most modern computers, servers, phones, and other IoT devices via industry leaders like Intel, AMD, and ARM [8].

References

1. Hirak, K., Hasin, A., Nazrul, H., Swarup, R., Dhruba, K.. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. https://arxiv.org/abs/1506.05101.

2. Holub, P. et al., Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health, Biopreservation and Biobanking. 16:2, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5906729/pdf/bio.2017.0110.pdf

3. Kondor, D., Hashemian, B., de Montjoye, Y., Ratti, C., Towards matching user mobility traces in large-scale datasets, IEEE Transactions on Big Data. Sept 2018. https://ieeexplore.ieee.org/document/8470173

4. Tramèr, F., Zhang, F., Juels, A., Reiter, M., Ristenpart, T., Stealing Machine Learning Models via Prediction APIs, Usenix, 2016. https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf

5. Bellwood, L., McCloud, S., Federated Learning: building better products with on-device data and privacy by default, Google AI, 2019. https://federated.withgoogle.com/

6. Bonawitz, K., et al. Towards Federated Learning at Scale: System Design. 2019. https://arxiv.org/abs/1902.01046

7. Dwork, C., Roth, A., The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science. 9:3. 2014.

8. Costan, V., Devadas, S., Intel SGX Explained. International Association for Cryptologic Research. 2016. https://eprint.iacr.org/2016/086.pdf