Developing Metrics for Privacy-Preserving Federated Learning

PEPR Conference – June 9, 2025

Curtis Mitchell, Emerging Technology Fellow @ xD

Agenda

1. Project Outline
2. Overview of Genomic Data and Privacy
3. Project Architecture
4. Red-Teaming Challenge & Example Results
5. Future Goals

1. Project Outline

A multi-agency effort to explore
privacy-preserving federated learning architectures.

NIST

United States Census Bureau

Project Partners

- NIST – Privacy Engineering Program & National Center of Cybersecurity Excellence (NCCoE)
- MITRE
- Knexus
- Census Bureau – xD

Project Outline

Problem space: sharing medical data across institutions and borders

Possible solution: privacy-preserving federated learning

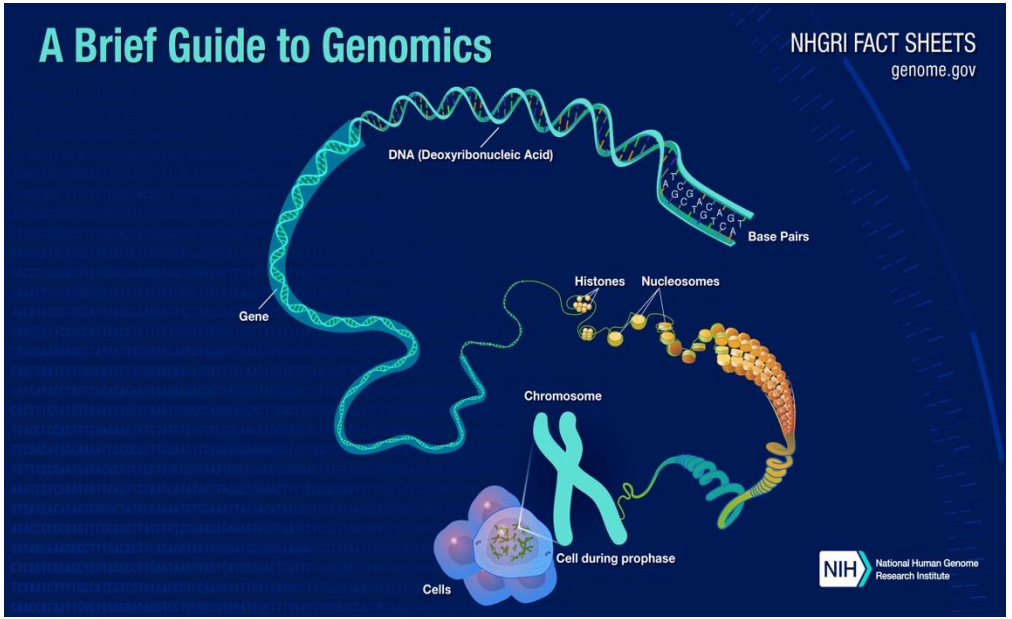
Validation: red-teaming of PPFL-created models and creating metrics on privacy-preservation

2. Genomic Data & Privacy

What is genomic data and
why is it important to protect?

Genomics Data

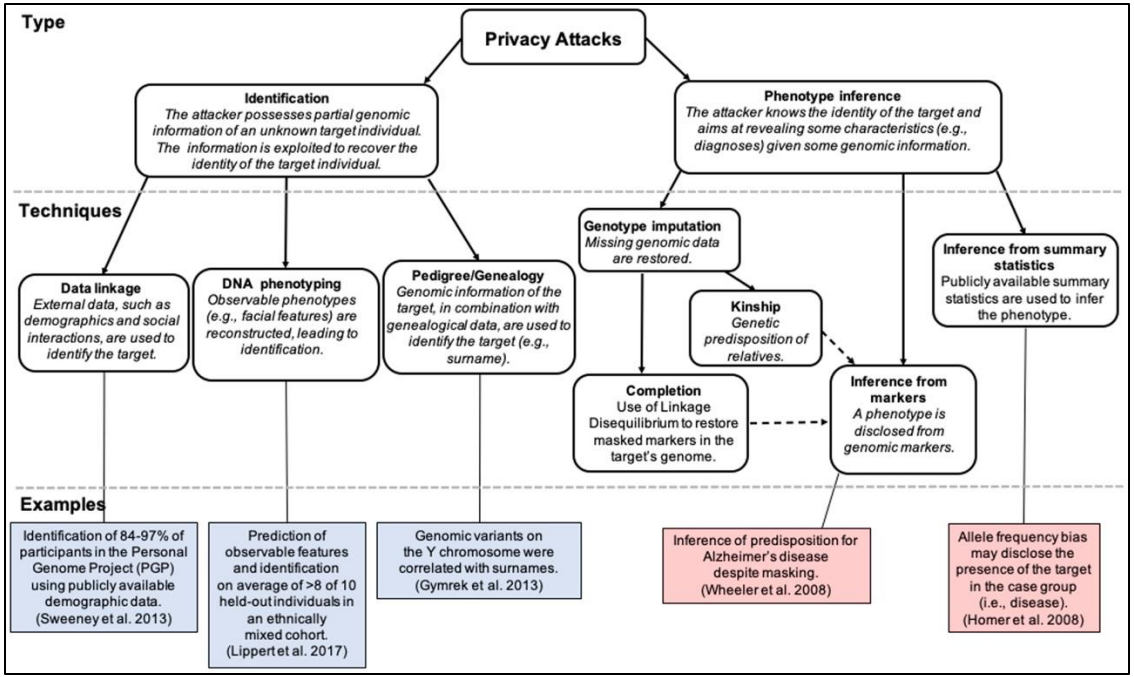
Genomics: collective characterization and quantification of all of an organism's genes, their interrelations and influence on the organism.



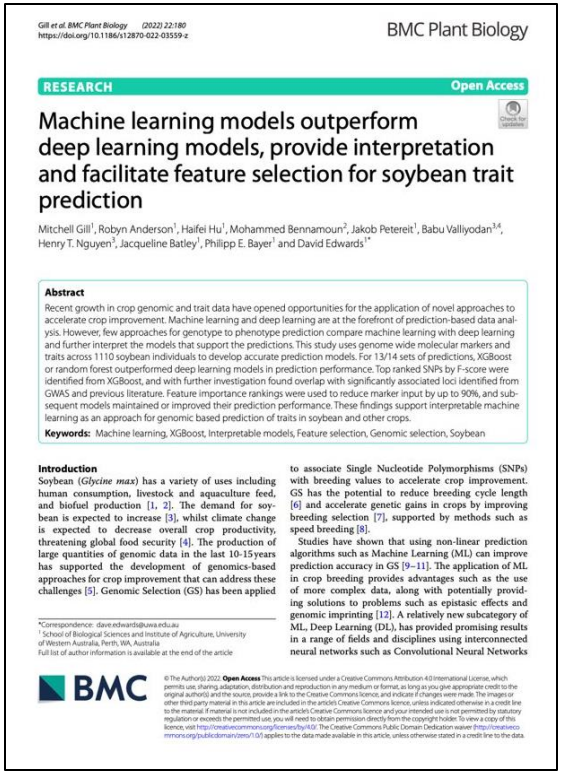
Genomics Data & Privacy

- "75 statistically independent SNPs* would suffice to uniquely identify an individual across the global population" - *Privacy Challenges and Research Opportunities for Genomic Data Sharing*, *Bonomi et al (2021)*

*SNP = Single Nucleotide Polymorphism



Genomics Data Source



Gill et al. BMC Plant Biology (2022) 22:180
https://doi.org/10.1186/s12870-022-03559-z

BMC Plant Biology

RESEARCH Open Access

Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction

Mitchell Gill[1], Robyn Anderson[1], Haifei Hu[1], Mohammed Bennamoun[2], Jakob Petereit[1], Babu Valliyodan[3,4], Henry T. Nguyen[3], Jacqueline Batley[1], Philipp E. Bayer[1] and David Edwards[1*]

Abstract

Recent growth in crop genomic and trait data have opened opportunities for the application of novel approaches to accelerate crop improvement. Machine learning and deep learning are at the forefront of prediction-based data analysis. However, few approaches for genotype to phenotype prediction compare machine learning with deep learning and further interpret the models that support the predictions. This study uses genome wide molecular markers and traits across 1110 soybean individuals to develop accurate prediction models. For 13/14 sets of predictions, XGBoost or random forest outperformed deep learning models in prediction performance. Top ranked SNPs by F-score were identified from XGBoost, and with further investigation found overlap with significantly associated loci identified from GWAS and previous literature. Feature importance rankings were used to reduce marker input by up to 90%, and subsequent models maintained or improved their prediction performance. These findings support interpretable machine learning as an approach for genomic based prediction of traits in soybean and other crops.

Keywords: Machine learning, XGBoost, Interpretable models, Feature selection, Genomic selection, Soybean

Introduction

Soybean (*Glycine max*) has a variety of uses including human consumption, livestock and aquaculture feed, and biofuel production [1, 2]. The demand for soybean is expected to increase [3], whilst climate change is expected to decrease overall crop productivity, threatening global food security [4]. The production of large quantities of genomic data in the last 10-15 years has supported the development of genomics-based approaches for crop improvement that can address these challenges [5]. Genomic Selection (GS) has been applied to associate Single Nucleotide Polymorphisms (SNPs) with breeding values to accelerate crop improvement. GS has the potential to reduce breeding cycle length [6] and accelerate genetic gains in crops by improving breeding selection [7], supported by methods such as speed breeding [8].

Studies have shown that using non-linear prediction algorithms such as Machine Learning (ML) can improve prediction accuracy in GS [9–11]. The application of ML in crop breeding provides advantages such as the use of more complex data, along with potentially providing solutions to problems such as epistasic effects and genomic imprinting [12]. A relatively new subcategory of ML, Deep Learning (DL), has provided promising results in a range of fields and disciplines using interconnected neural networks such as Convolutional Neural Networks

*Correspondence: dave.edwards@uwa.edu.au
[1] School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia
Full list of author information is available at the end of the article

BMC

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction, BMC Plant Biology 2022.

Genomics Data Format

- Columns = organisms
- Rows = accessions (specific location of base pairs on a gene)
- Prediction is categorical or continuous value

(here, value = **seed oil percentage)

	AB-02	BR-24	ESS	HN002	HN003	HN004	HN005	HN007
Gm01_20970	T/T	T/T	T/T	T/T	T/T	C/C	T/T	C/C
Gm01_21336	C/C	G/G	G/G	C/C	C/C	G/G	C/C	G/G
Gm01_26249	T/T	A/A	A/A	T/T	T/T	T/T	T/T	T/T
Gm01_26286	G/G	G/G	G/G	G/G	G/G	T/T	G/G	T/T
Gm01_39069	T/T	C/C	C/C	T/T	T/T	T/T	T/T	T/T
Gm01_39178	G/G	A/A	A/A	G/G	G/G	G/G	G/G	G/G
Gm01_39297	T/T	A/A	A/A	T/T	T/T	T/T	T/T	T/T
Gm01_39560	A/A	G/G	G/G	A/A	A/A	A/A	A/A	A/A
Gm01_39592	A/C	A/A	A/A	C/C	C/C	A/A	C/C	A/A
...	...	...	...	...	...	...	...	...
Value	16.8	20.6	20.9	18.5	17.5	18.9	15.5	15.1

3. Project Architecture

Privacy preserving federated learning.

Project Libraries

CNNs with PyTorch

DP-CNNs

with PyTorch + Opacus

FL with Flower

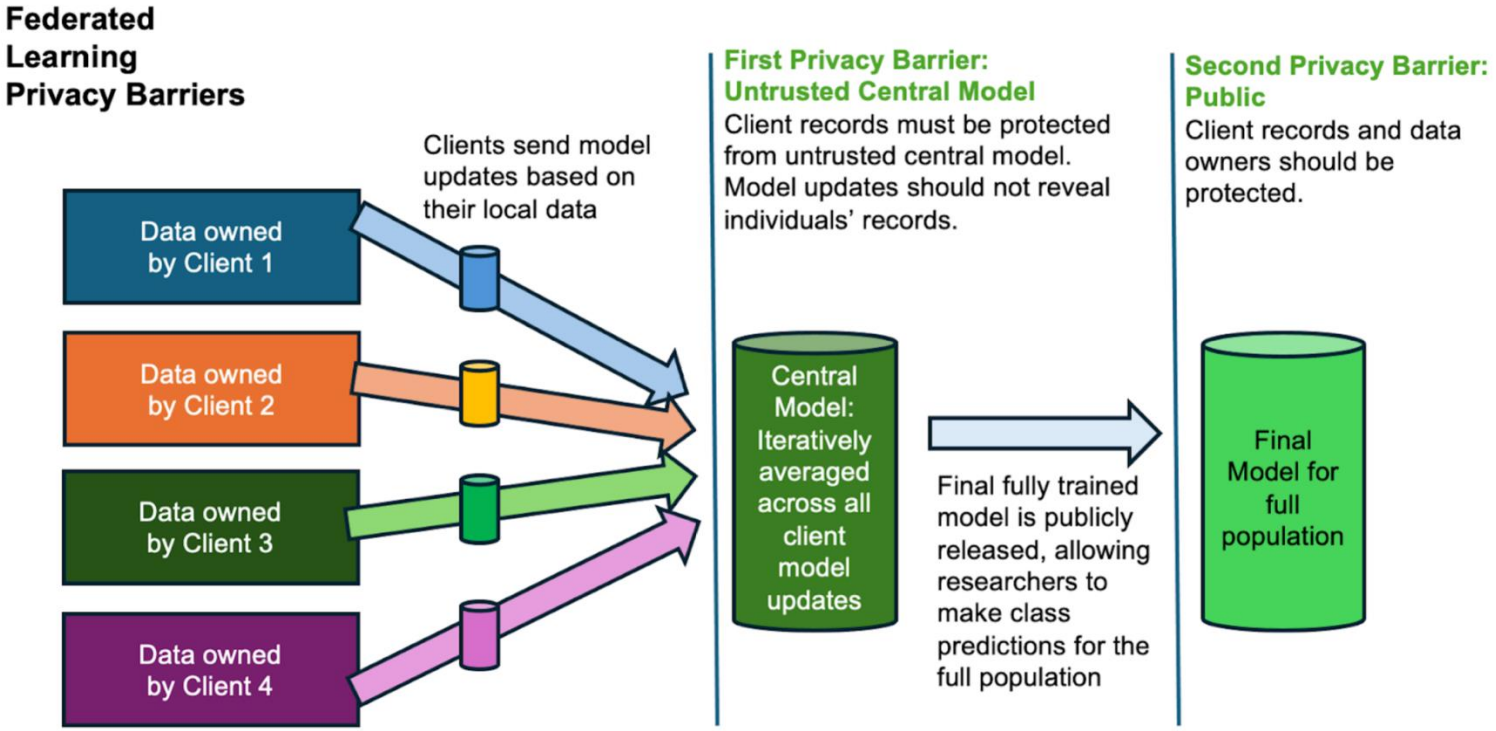
Using FedAvg algorithm

Project Architecture & Parameters

Comparison of performance and privacy across different scenarios using federated learning with 4 clients for 100 epochs:

- Regular CNNs
- DP-CNNs with high privacy ($\varepsilon = 10$, clipping = 2.0)
- DP-CNNs with low privacy ($\varepsilon = 200$, clipping = 2.0)

Project Trust Model



4. Red-Teaming Challenge & example results

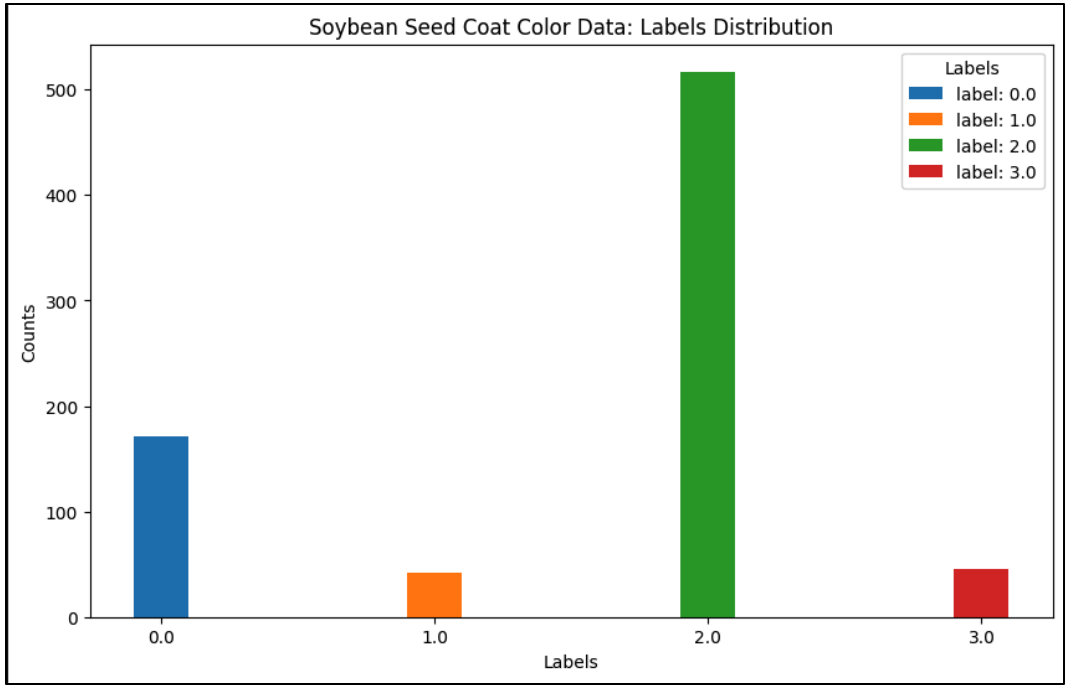
Breaking privacy using
membership inference attacks.

Questions Posed for the Challenge

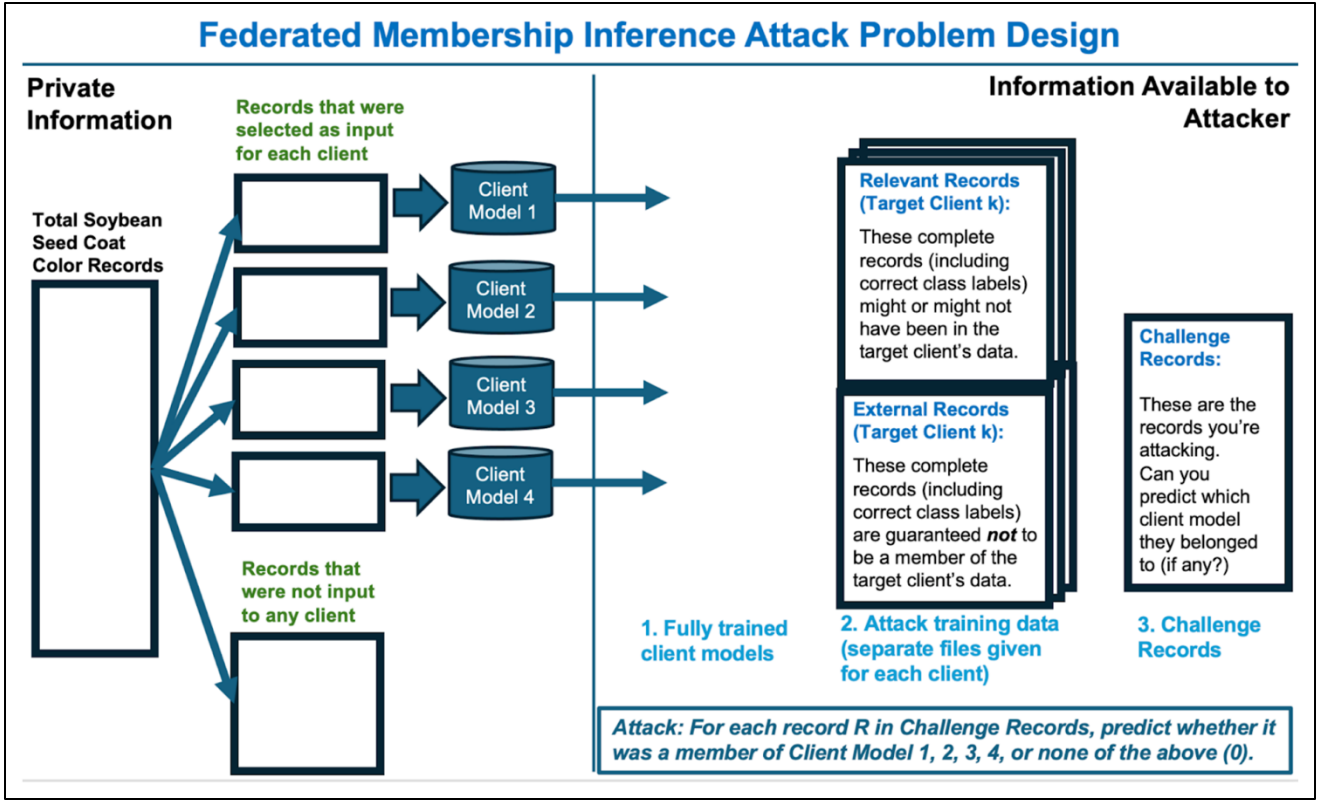
- Are the non-private CNNs vulnerable to membership inference attacks?
- Can differential privacy help protect client data?
- Can it help even if epsilon is so large that the added privacy noise doesn't affect the overall utility?
- Different clients have different local data distributions – does the effectiveness of differential privacy differ for different clients?

Don't Spill The Beans

- Red-teaming exercises to test privacy protections of various architectures
- Phase 1 is a membership inference attack
- Includes several tasks such as label predictions of both categorical and continuous values
- Participants are provided a mix of relevant data (mimicking training data) and external records (same format but not in training data)



Red-Teaming Problem Design



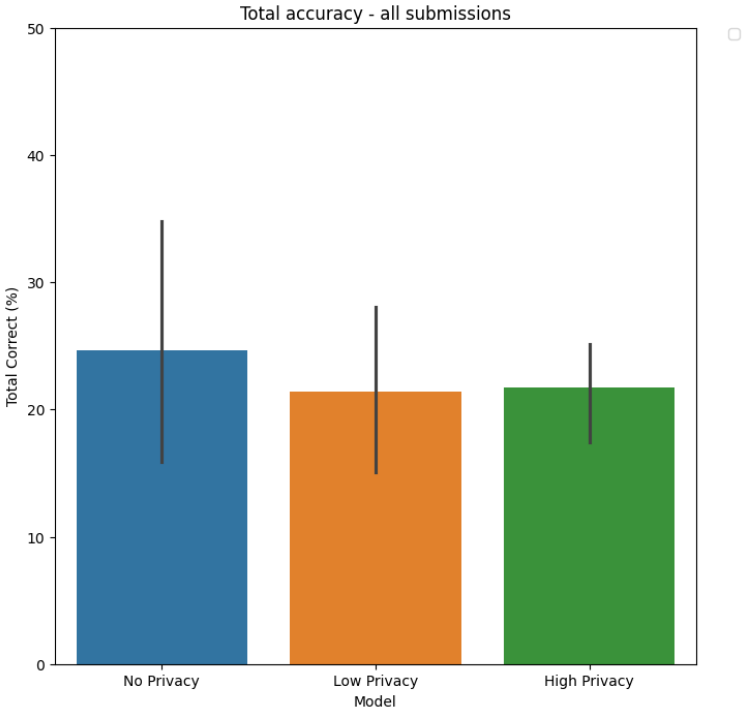
Phase 1 Results - Leaderboard

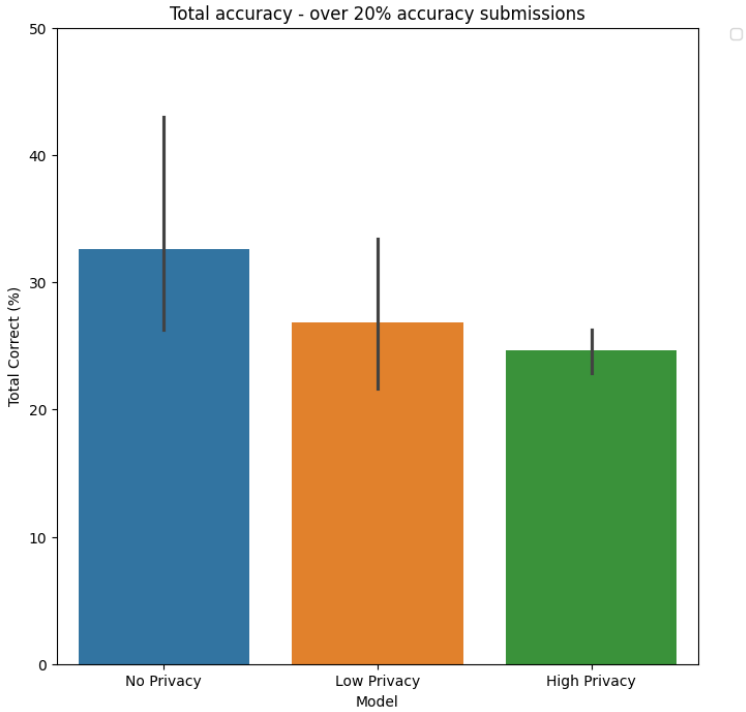
- Most teams manage to re-identify 20-30% of records, even for DP-based models

Leaderboard

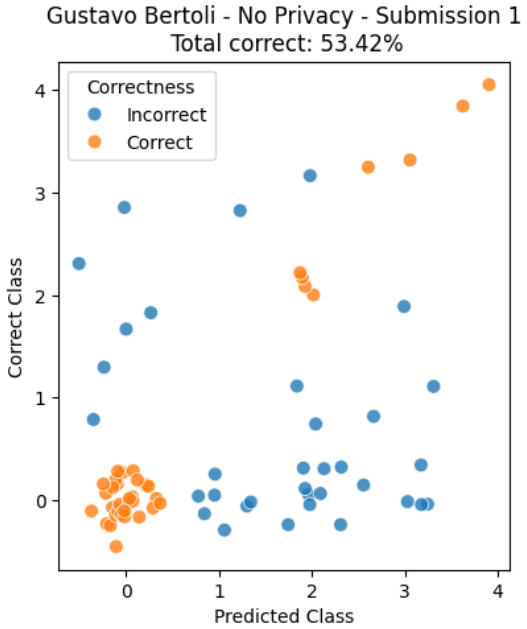
Rank	Team	Submission #	Percent Accuracy on Challenge Records			
			CNN	Low Privacy (DP 200)	High Privacy (DP 10)	Simple Sum Total
1	Gustavo Bertoli	1	53.42	38.36	24.66	116.44
1	Gustavo Bertoli	2	30.14	20.55	26.03	76.72
2	MITRE PPFL	1	27.4	26.03	20.55	73.98
2	MITRE PPFL	2	26.03	28.77	26.03	80.83
3	Arrakus	1	26.03	19.18	27.4	72.61
3	Arrakus	2	10.96	17.81	20.55	49.32
4	Data Preservers	1	17.81	23.29	13.7	54.8
5	Beam Me Up Scotty	1	5.48	4.11	8.22	17.81

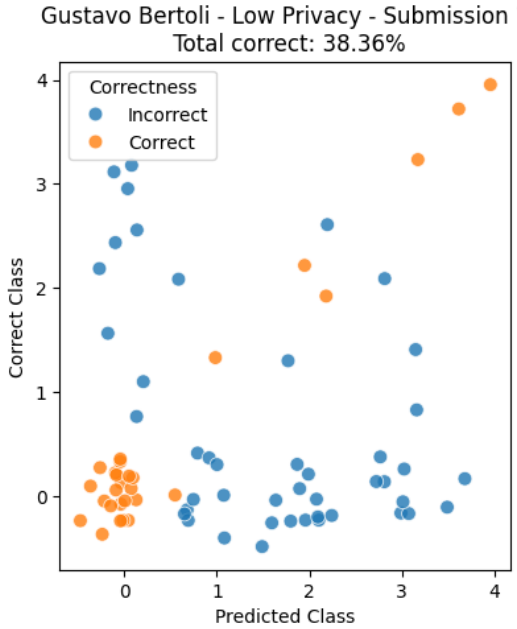
Phase 1 Results – Total % Correct

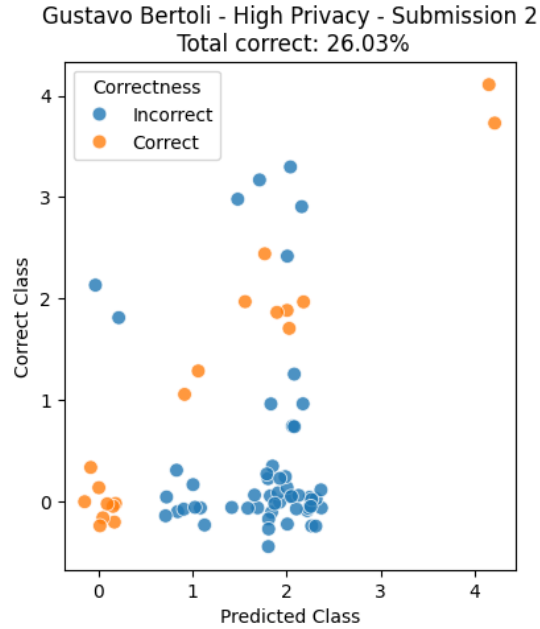




Phase 1 Results – Best Submission







Example Strategy from MITRE PPFL Team

- Theorized that soybeans would have similar genomes by region
- Used PCA to reduce dimensionality and ease classifier training
- Compared performance of multiple classifiers using PyCaret
- Classifiers were trained on all data (relevant and external)

2	MITRE PPFL	1	27.4	26.03	20.55	73.98
2	MITRE PPFL	2	26.03	28.77	26.03	**80.83**

5. Future Goals

More red-teaming and
defining usable privacy metrics

Future Ideas - Data

- Proof-of-concept with plant genomic data - DONE
- Animal cancer data
- Human cancer data

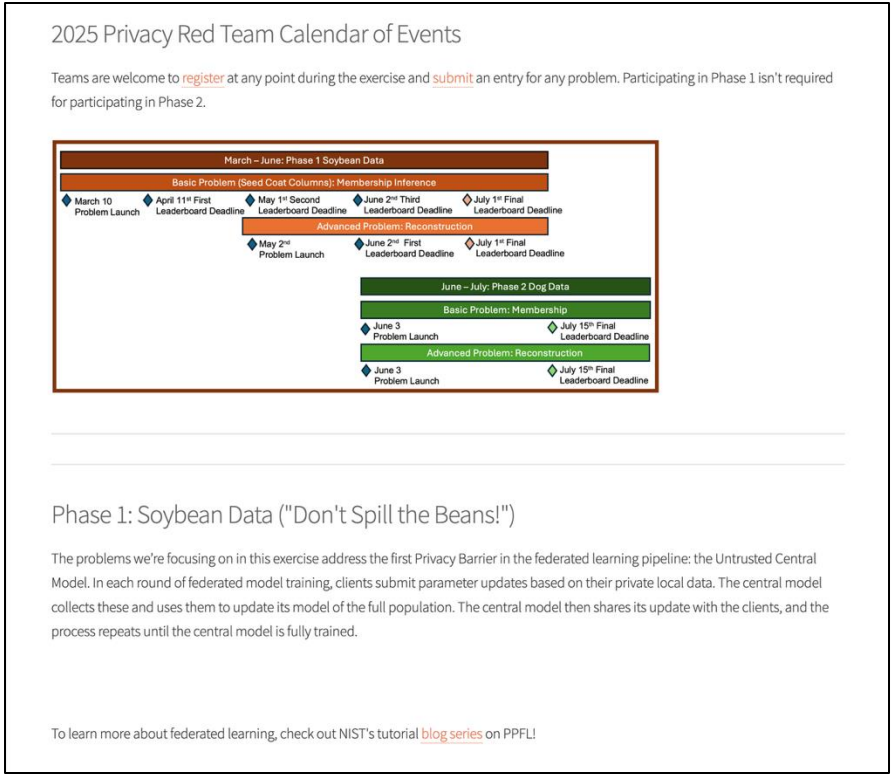
Potential Outcomes

Building a PPFL system for cancer research with robust privacy could enable:

- Better understanding of how ML/FL impact privacy
- More impactful healthcare research
- Better health outcomes and privacy protections for patients

Join the Red-Teaming Challenge

- Phase 2 challenge (reconstruction attacks) runs from May-July
- https://pages.nist.gov/genomics_ppfl



Acknowledgements

Gary Howarth & Justin Wagner (NIST)

Christine Task & Karan Bhagat (Knexus)

Amy Hilla & Rebecca Steinberg (MITRE)

Jess Stahl (Census Bureau / xD)

Q & A

Appendix

Initial Results - accuracy