National Institute of General Medical Sciences

UTHealth | School of Biomedical Informatics
The University of Texas Health Science Center at Houston

The iDASH Competition : Progress of Homomorphic Encryption for Genomic Privacy

Miran Kim

University of Texas, Health Science Center at Houston

Lattices: From Theory to Practice, Simons Workshop, April 30

Supported by NHGRI R13HG009072

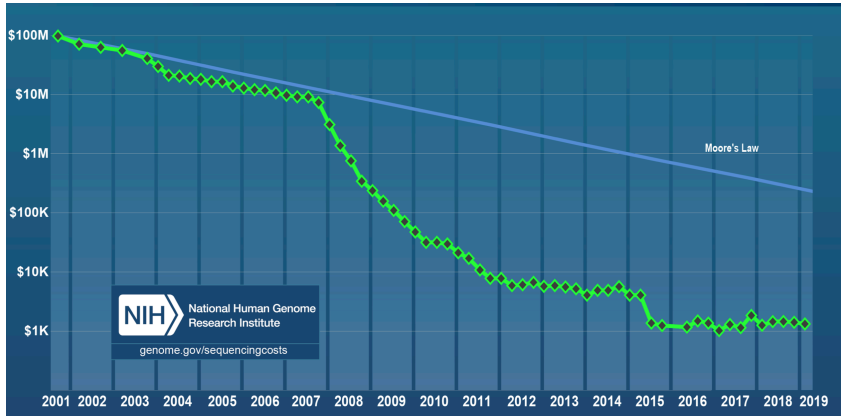
Joint work with Dr. Xiaoqian Jiang (UTHealth), Arif Harmanci (UTHealth), Haixu Tang (IUB), XiaoFeng Wang (IUB), Lucila Ohno-Machado (UCSD), Tsung-Ting Kuo (UCSD),

Genome Revolution

- Whole Genome Sequencing is getting cheaper.
  - 2000: $3 billion
  - 2014: $1,000

- Data sharing is very important for biomedicine to speedup discovery and promote research.

- NIH Genomic Data Sharing policy allows the use of cloud computing services for storage and analysis of controlled-access data (2014).

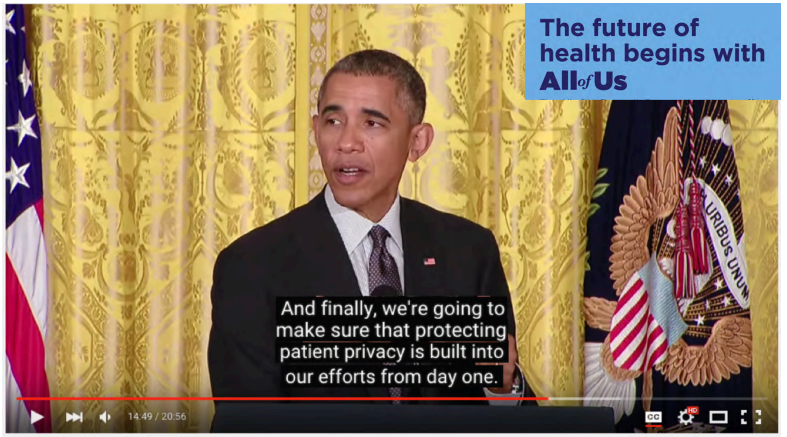


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Human Genomic Data Sharing

- But genomic data are highly sensitive.
  - Re-identification leak privacy
    - Lin et al. (2004 Science): SNPs ≥ 75 can identify a single person
    - Gymrek et al. (2013 Science): surnames can be recovered from personal genomes
  - Genetic discrimination, Genetic disease disclosure
  - A great fear of unknown

- Privacy Protection Law
  - HIPPA (US): Health information regulation law, de-identification
  - GDPR (EU): General Data Protection Regulation

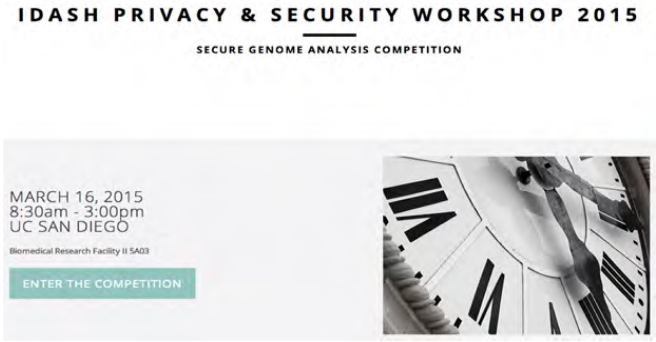


President Obama Speaks on the Precision Medicine Initiative

Community Effort in Promoting Genomic Privacy

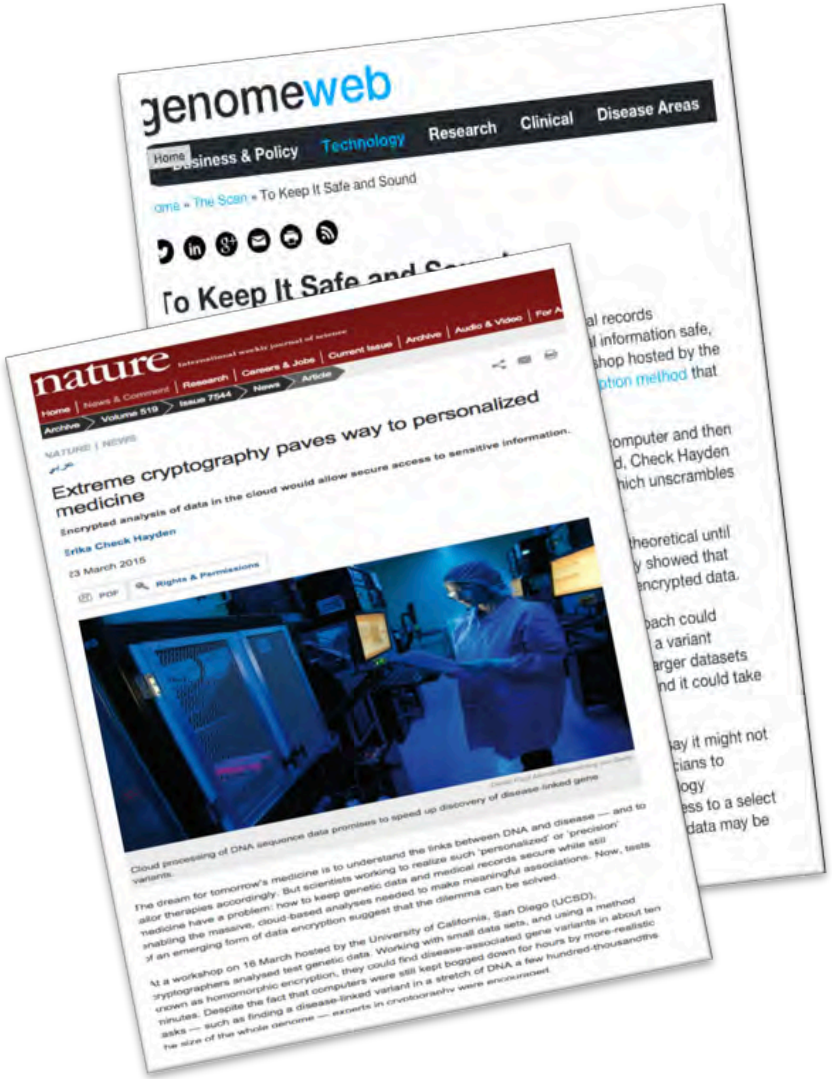
2014-2019 iDASH genomic data privacy and security protection competition http://www.humangenomeprivacy.com

Sponsored by NIH, Human Longevity, Genecloud, Baidu, illumine, PlatON



iDASH Privacy Workshop

- Motivated by real-world biomedical challenges and with participation of crypto experts and biomedical researchers
- Developed practical yet rigorous solutions for privacy preserving genomic data analysis
- Demonstrate the feasibility of secure genome data analysis using HE, differential privacy, multi-party computation, SGX
- Reported in the media (iDASH'15)
  - e.g., Nature News, GenomeWeb, Donga, Microsoft Research News



http://www.nature.com/news/extremecryptography-paves-way-to-personalizedmedicine-1.17174

Summary of Challenges and Tasks

2015	12 Teams	• HE-based genome analysis (DNA sequence comparison) • MPC-based genome analysis
2016	50 Teams	• Testing for genetic disease on homomorphically encrypted genomes • MPC-based privacy-preserving search of similar cancer patients across organizations • Protecting queries in Beacon service
2017	65 Teams	• Homomorphic Logistic Regression Training • Secure record de-duplication • Secure GWAS using SGX
2018	64 Teams	• Secure Parallel Genome Wide Association Studies using Homomorphic Encryption • Blockchain-based immutable logging and querying for cross-site genomic dataset access audit trail • MPC-based secure search of DNA segments in large genome databases
2019	105 Teams	• Secure Genotype Imputation using Homomorphic Encryption (54 teams) • Distributed Gene-Drug Interaction Data Sharing based on Blockchain and Smart Contracts • Privacy-preserving Machine Learning as a Service on SGX • MPC-based Secure Collaborative Training of Machine Learning Model

iDASH'15&16

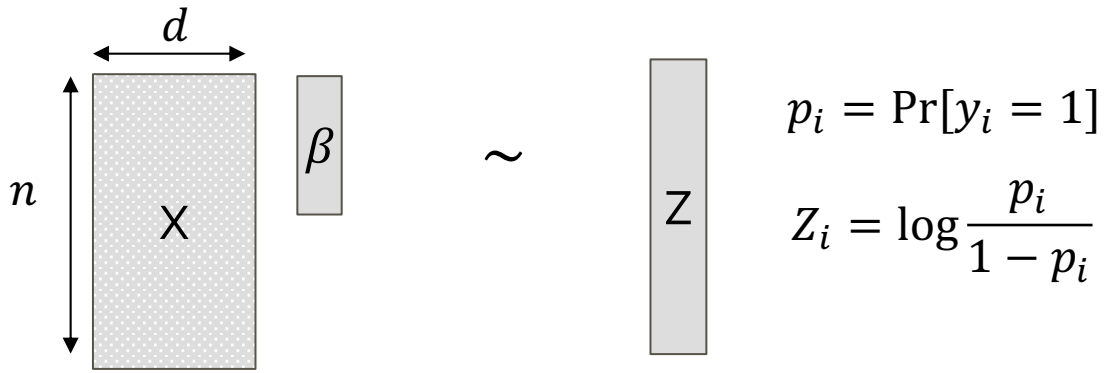
- p: plaintext modulus
  - [$p =$ Prime]: $equal(a, b) = 1 - (a - b)^{p-1} \in \mathbb{Z}_p$
  - [$p = 2$]: $equal(a, b) = 1 \oplus a \oplus b \in \mathbb{Z}_2$ where a, b: one-bit.

  So, $equal(a, b) = \prod_{i=1}^{l}(1 \oplus a_i \oplus b_i)$ when a, b: l-bits.

Year	Task	Scheme	Dataset	Time	Memory
2015	Hamming Distance	BGV	100K sequences	8 min	2.2 GB
	Approximate Edit Distance	BGV	10K sequences	3 min	1.3 GB
2016	Genetic testing	BFV	1 query (1 genetic variant) in 50 VCF files (100K)	1 min	83 MB

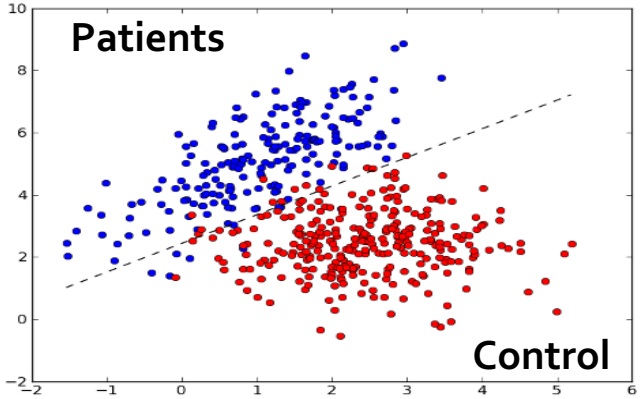
iDASH'17: Logistic Regression Training (1)

- Build a machine learning model to predict a disease
  - Given phenotype $y_i \in \{\pm 1\}$, genotype $X_i \in \mathbb{R}^d$,
  - Goal: find $\beta \in \mathbb{R}^d$ s.t.



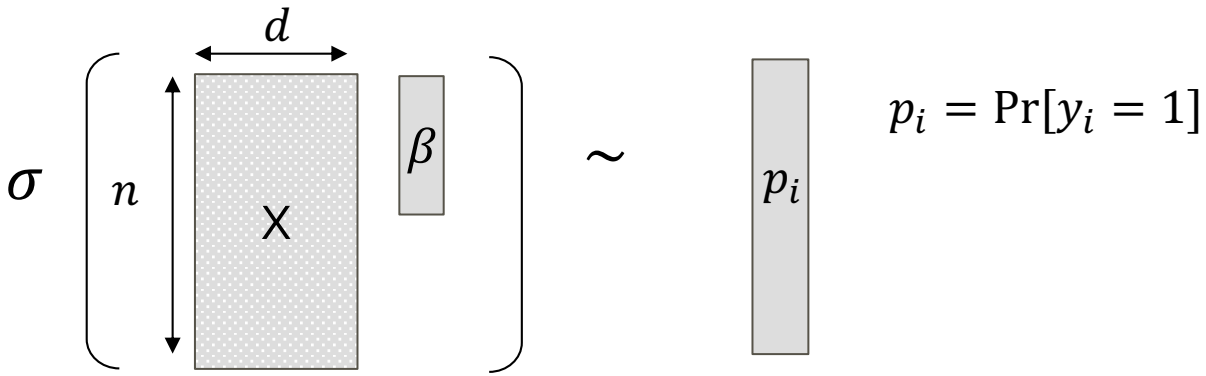
$$p_i = \Pr[y_i = 1]$$

$$Z_i = \log \frac{p_i}{1 - p_i}$$

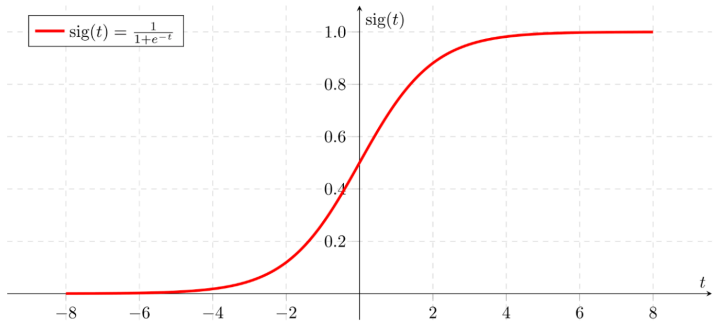


iDASH'17: Logistic Regression Training (1)

- Build a machine learning model to predict a disease
  - Given phenotype $y_i \in \{\pm 1\}$, genotype $X_i \in \mathbb{R}^d$,
  - Goal: find $\beta \in \mathbb{R}^d$ s.t.



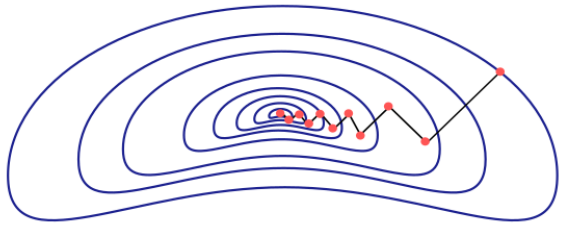
$$p_i = \Pr[y_i = 1]$$



logistic: $\sigma(x) = 1/(1 + e^{-x})$

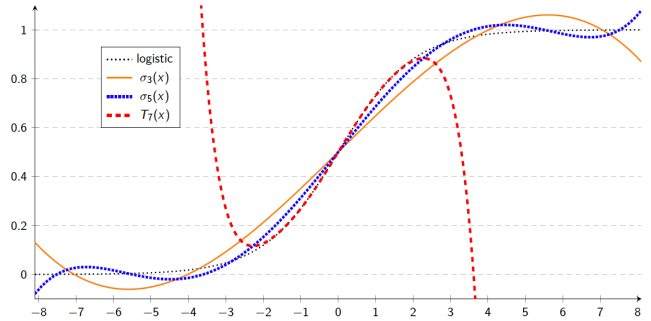
- Machine-learning approach
  - Problem: Minimize the loss $J(\beta) = -\sum_i \log\ [\ 1 + \exp(\ -y_i X_i \beta)\ ]$
  - Gradient decent method : Update $\beta \leftarrow \beta - \alpha \cdot \nabla_\beta J$

    where $p = (p_i) = \sigma(X_i\beta),\ \nabla_\beta J = X^T(y - \mathrm{p})$
  - Newton method: $\alpha = (\nabla^2_\beta J)^{-1}$ where $\nabla^2_\beta J = -X^T \cdot \mathrm{diag}\big(\mathrm{p}(1-\mathrm{p})\big) \cdot X$



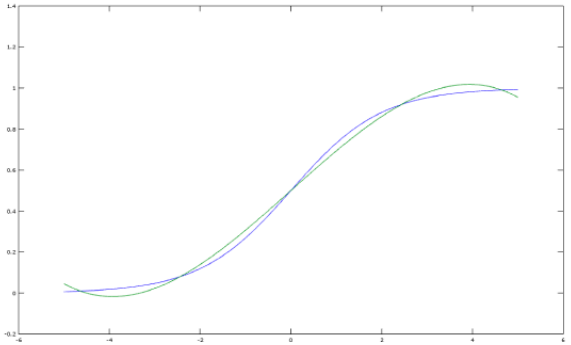
iDASH'17: Logistic Regression Training (2)

- Challenge
  - Non-polynomial functions
  - Real number arithmetic
  - Training algorithms are recursive: depth = O(num iteration)



Least-squared approximation

- Polynomial Approximation
  - Taylor series expansion
  - Bernstein expansion: $B_n(f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_{k,n}(x)$ where $b_{k,n}(x) = \binom{n}{k} x^k (1-x)^{n-k}$
  - Least-squared approximation: minimize $\frac{1}{|I|} \int \int_I \ (f(x) - g(x))^2 dx$
  - Minimax approximation: minimize $\inf\{\|f(x) - g(x) : x \in I\|\}$


Minimax approximation

iDASH'17: Logistic Regression Training (3)

- 1422 records + 18 features for training/ 80-bit security

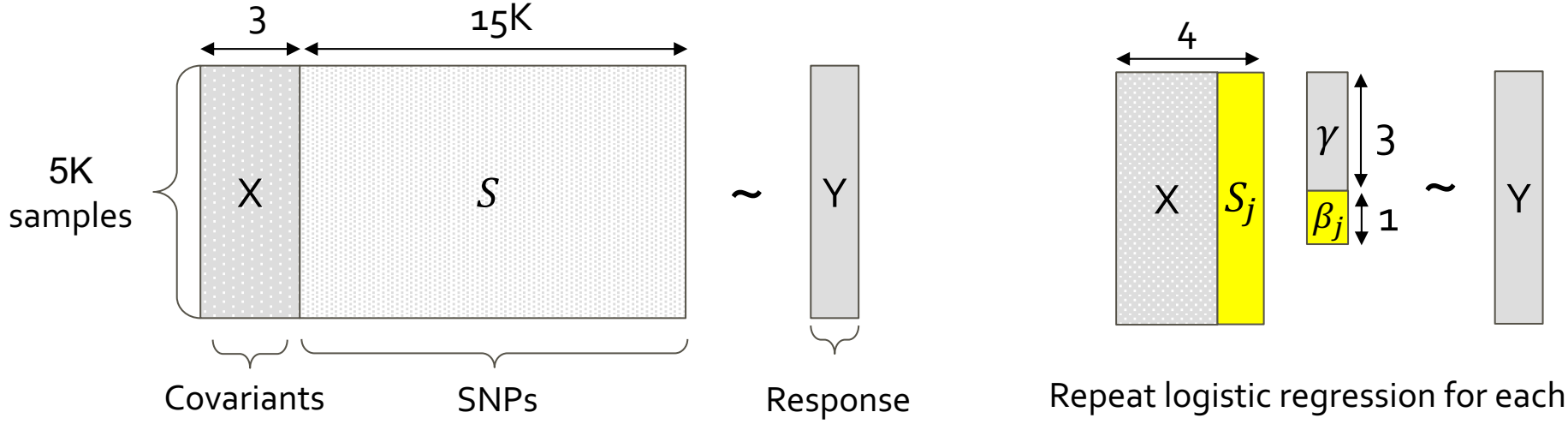
Team	Scheme	Approach	Encryption		Secure learning		Decryption		Total time (min)	AUC (0.7136)
			Time (min)	Size (MB)	Time (min)	Size (MB)	Time (min)	Size (MB)		
CEA LIST	-	-	1.30	53	2206	238	0.003	0.350	2207	0.6930
EPFL	BFV	Bernstein expansion 1-iter. of GD	1.63	1011	15	1498	0.017	7	17	0.6584
KU Leuven	BFV	Taylor series expansion 1-iter. of Newton's method $(\nabla^2 L(\beta) \sim \frac{1}{4} X^T X)$	4.30	4904	155	7266	0.913	10	161	0.6722
Microsoft	BFV	Minimax approximation 17-iters of GD Homomorphic floor ($\lfloor m/p \rfloor$) inside bootstrapping	11.34	1945	385	26299	0.033	76	396	0.6574
Saarland	-	-	1.63	65536	48	29752	7.355	65536	57	N/A
SNU& UCSD	CKKS	Least-squared approximation 7-iters of Nesterov's GD Built-in rescaling in CKKS	0.06	537	10	2775	0.050	64	**10**	**0.6934**

iDASH'18: Genome Wide Association Studies (GWAS)

- Task: test the associations between genotypes and phenotypes
  - Given phenotype $y_i \in \{\pm 1\}$, covariate $X_i \in \mathbb{R}^d$, genotype $s_{ij} \in \{0,1\}$,

    find $\beta_j \in \mathbb{R}$ s.t. $\Pr[y_i = 1] = \sigma(X_i\gamma + S_j\beta)$.
  - 5K samples * (3 covariates : age, weight, height + 15K SNPs)
  - Naïve solution: repeat logistic regression model training 15K times

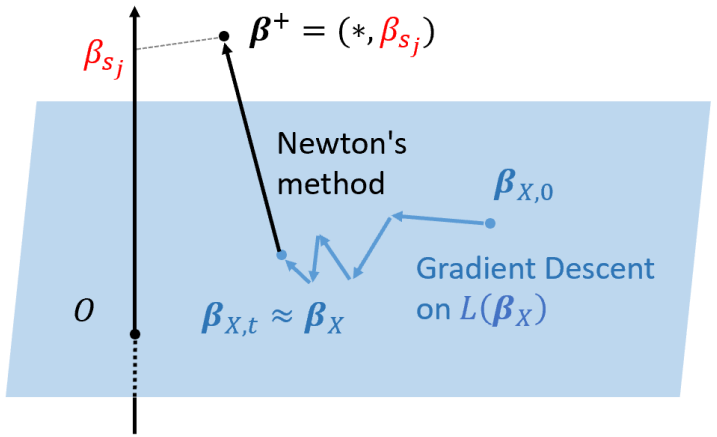
    (one SNP at a time is too costly)



iDASH'18: Genome Wide Association Studies (GWAS)

- Main Idea : Semi-parallel logistic regression [SLGE13]
  - Assume the parameters for covariants will stay nearly the same for all SNPs.
  - Strategy
    - Step-1: Pre-train a model on covariates X (only one time)
    - Step-2: Parallelize regression on all SNPs (one-step of Newton's method)
    - Challenge: $\left(\nabla^2_\beta J\right)^{-1} = (-X^T \cdot W \cdot X)^{-1}$

      where $p = \sigma(X^T\beta)$, $\mathrm{W} = \mathrm{diag}\left(p(1-p)\right)$



[SLGE13] Sikorska K, Lesaffre E, Groenen PFJ, Eilers PHC. GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. BMC Bioinformatics 2013 May 28;14:166. PMID:23711206

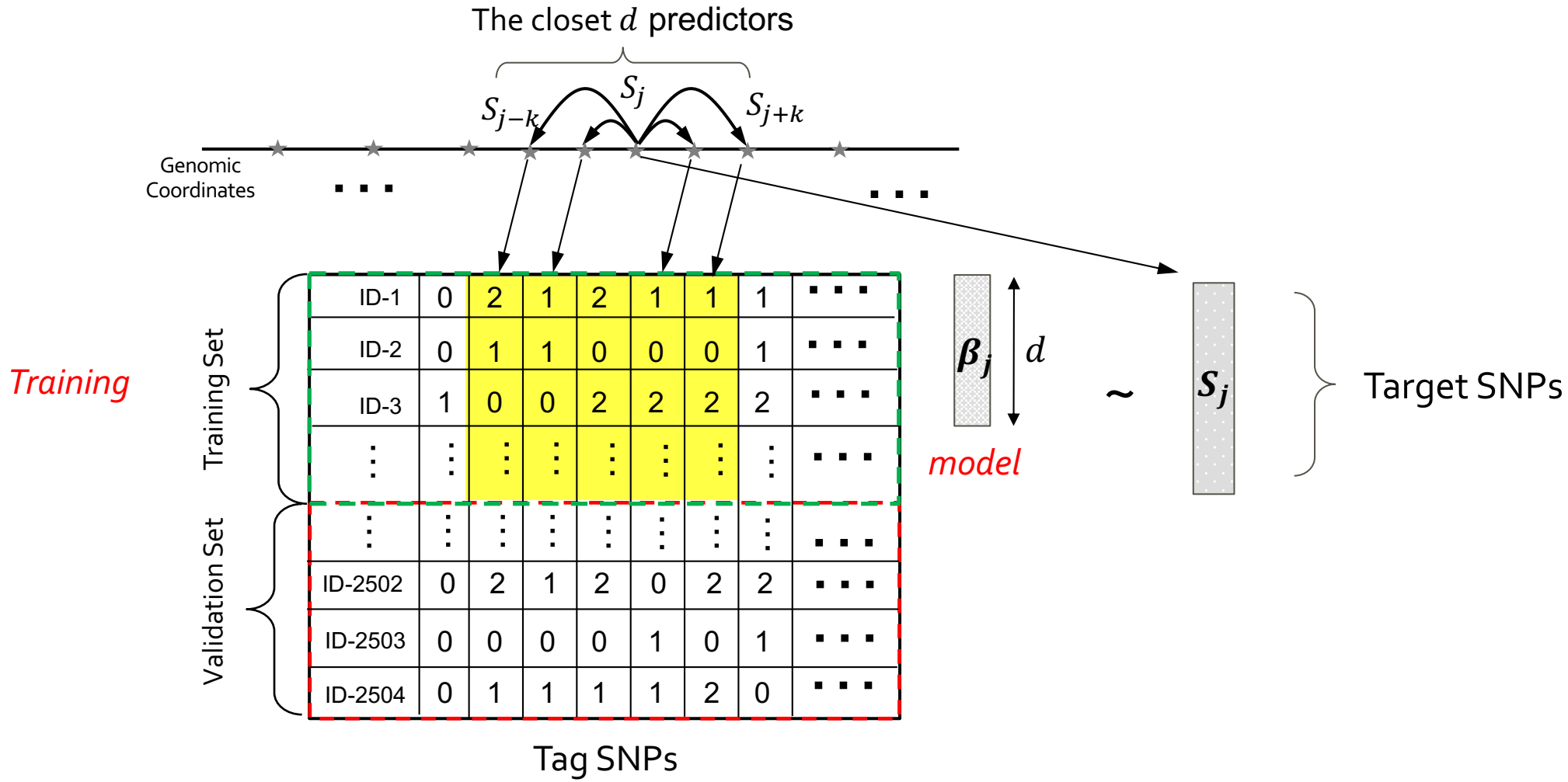
iDASH'18: Genome Wide Association Studies (GWAS)

- 5K samples * (3 covariates : age, weight, height + 15K SNPs)

Team	Scheme	Approach	Enc-to-end performance		Evaluation result (F1- Score)			
					1E-2		1E-5	
			Time	Memory	Gold	Semi	Gold	Semi
A*FHE	CKKS	Encrypt $(X^T X)^{-1}$ and compute W^{-1} using Newton method	15.3 h	3.7 GB	0.977	0.999	0.966	0.998
Chimera	TFHE & CKKS	LogReg: Gate bootstrapping + sigmoid evaluation $X^T W X \sim \frac{1}{4}\mathrm{Id}$ (assuming X: orthogonal)	3.3 h	10.1 GB	0.979	0.993	0.982	0.974
Delft Blue	CKKS	-	31 h	10.6 GB	0.965	0.969	0.884	0.849
Duality Inc	RNS-CKKS	Chebyshev approximation of the sigmoid Adjugate & determinant	**3.8 min**	10.0 GB	**0.982**	**0.993**	**0.990**	**0.973**
IBM	CKKS	CKKS-complex	23 min	14.8 GB	0.913	0.911	0.053	0.06
SNU	CKKS	Adjugate & determinant	52 min	14.8 GB	0.975	0.984	0.932	0.905
UCSD	RNS-CKKS	Adjugate & determinant	**1.7 min**	14.5 GB	**0.983**	**0.993**	**0.995**	**0.967**

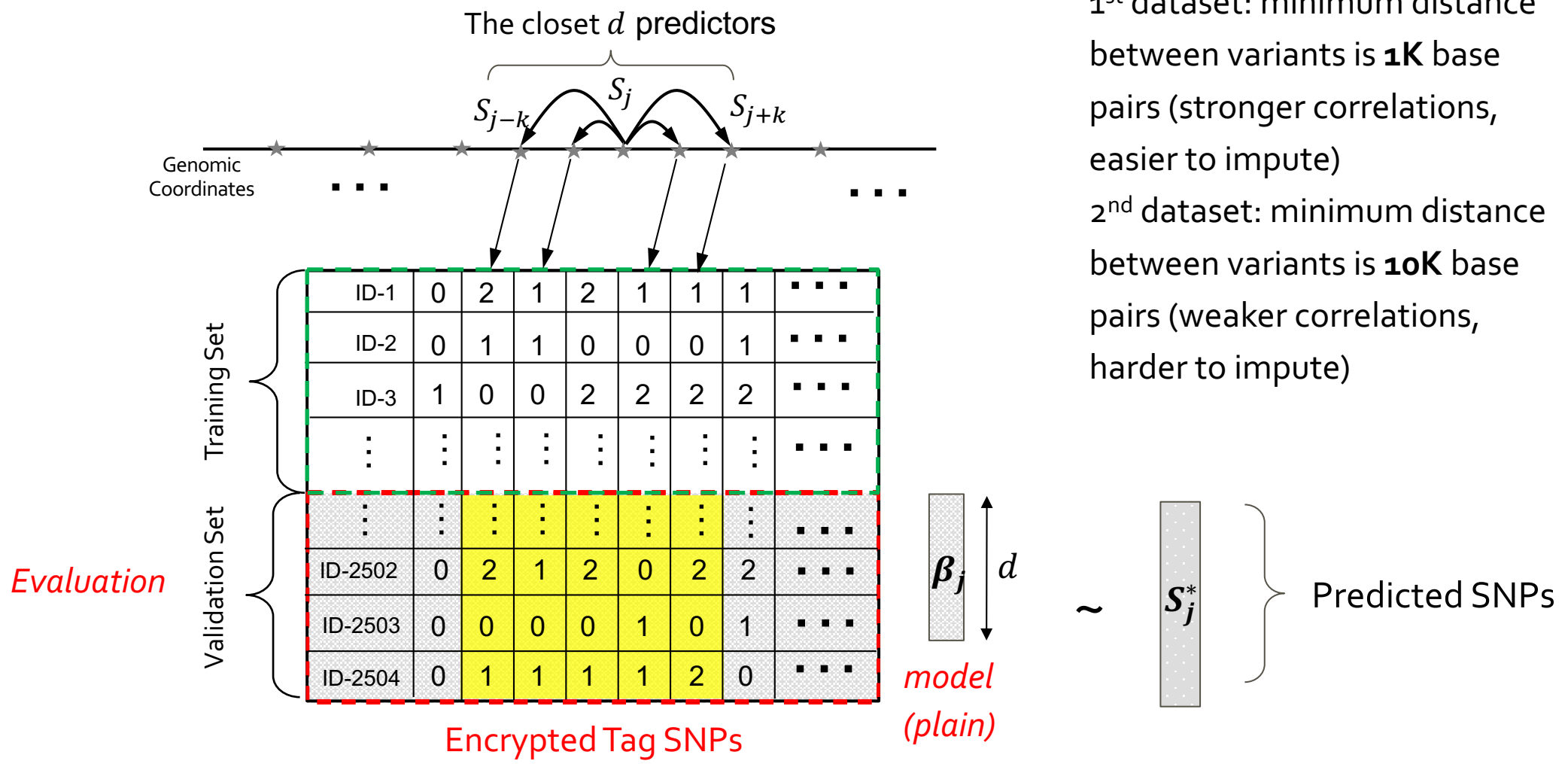
iDASH'19: Genotype Imputation

- Estimate the missing genotype



iDASH'19: Genotype Imputation

- Estimate the missing genotype



1st dataset: minimum distance between variants is **1K** base pairs (stronger correlations, easier to impute)

2nd dataset: minimum distance between variants is **10K** base pairs (weaker correlations, harder to impute)

iDASH'19: Genotype Imputation

- Tag-SNPs for testing: 250 samples * 9500 SNPs
- Target-SNPs for testing: 250 sample * 500 SNPs

Teams	Scheme	Approach	Performance					
			1K			10K		
			Time	Memory	Accuracy	Time	Memory	Accuracy
A*FHE	CKKS	Logistic regression	8.5 min	1.1 GB	0.9956	4.6 min	0.6 GB	0.9609
CodeHopper/ Temple	CKKS	Convolutional neural networks (2 conv + 1 fc)	9.1 min	2.7 GB	0.9959	7.6 min	2.7 GB	0.9435
EPFL	CKKS	Logistic regression (d=32)	22 sec	4.3 GB	0.9936	23 sec	4.3 GB	0.9705
Gerstin-MoMA Labs (Yale)	BFV	-	2.1 min	0.8 GB	0.9803	2 min	0.9 GB	0.9521
SNU	CKKS	1-hidden layer neural network (d=101,38)	2.6 min	0.9 GB	0.9966	51 sec	0.6 GB	0.9750
TFHE-Chimera	TFHE & CKKS	Logistic regression (d=50, 35) Coefficient packing strategy	3.5 sec	0.2 GB	0.9971	0.8 sec	0.03 GB	0.9763

References

- M. Kim and K. Lauter, "Private genome analysis through homomorphic encryption". BMC Med Inform Decis Mak. 2015.
- GS. Cetin, H. Chen, K. Laine, K. Lauter, P. Rindal, and Y. Xia, "Private Queries on Encrypted Genomic Data". BMC Med Genomics. 2017.
- A. Kim, Y. Song, M. Kim, K. Lee, and J.H. Cheon. "Logistic Regression Model Training based on the Approximate Homomorphic Encryption". BMC Med Genomics. 2018.
- C. Bonte and F. Vercauteren. "Privacy-Preserving Logistic Regression Training". BMC Med Genomics. 2018.
- H. Chen. R. Gilad-Bachrach, K. Han, Z. Huang, A. Jalali, K. Laine, and K. Lauter, "Logistic regression over encrypted data from fully homomorphic encryption". BMC Med Genomics. 2018.
- M. Kim, Y. Song, B. Li, and D. Micciancio, "Semi-parallel Logistic Regression for GWAS on Encrypted Data". To appear in BMC Med. Genomics.
- M. Blatt, A. Gusev, Y. Polyakov1, K. Rohlo, and V. Vaikuntanathan, "Optimized Homomorphic Encryption Solution for Secure Genome-Wide Association Studies". To appear in BMC Med. Genomics.
- S. Carpov, N. Gama, M. Georgieva, and J. Ramon Troncoso-Pastoriza2, "Privacy-preserving semi-parallel logistic regression training with Fully Homomorphic Encryption". To appear in BMC Med. Genomics.
- D. Kim, Y. Son, D. Kim, A. Kim, S. Hong, and J. H. Cheon, "Privacy-preserving Approximate GWAS computation based on Homomorphic Encryption". To appear in BMC Med. Genomics.
- J. J. Sim, F. M. Chan, S. Chen, B. H. M. Tan, and K. M. M. Aung, "Achieving GWAS with Homomorphic Encryption". To appear in BMC Med. Genomics.