

<https://github.com/shilab/Genetic-Privacy-and-Security/>

# **Genetic Privacy: Risks, Ethics, Regulations, and Protection Techniques**

Xinghua Mindy Shi

Lehigh University

April 4, 2023

Email: [mindyshi@temple.edu](mailto:mindyshi@temple.edu)    <https://cis.temple.edu/~mindyshi/>

# Outline

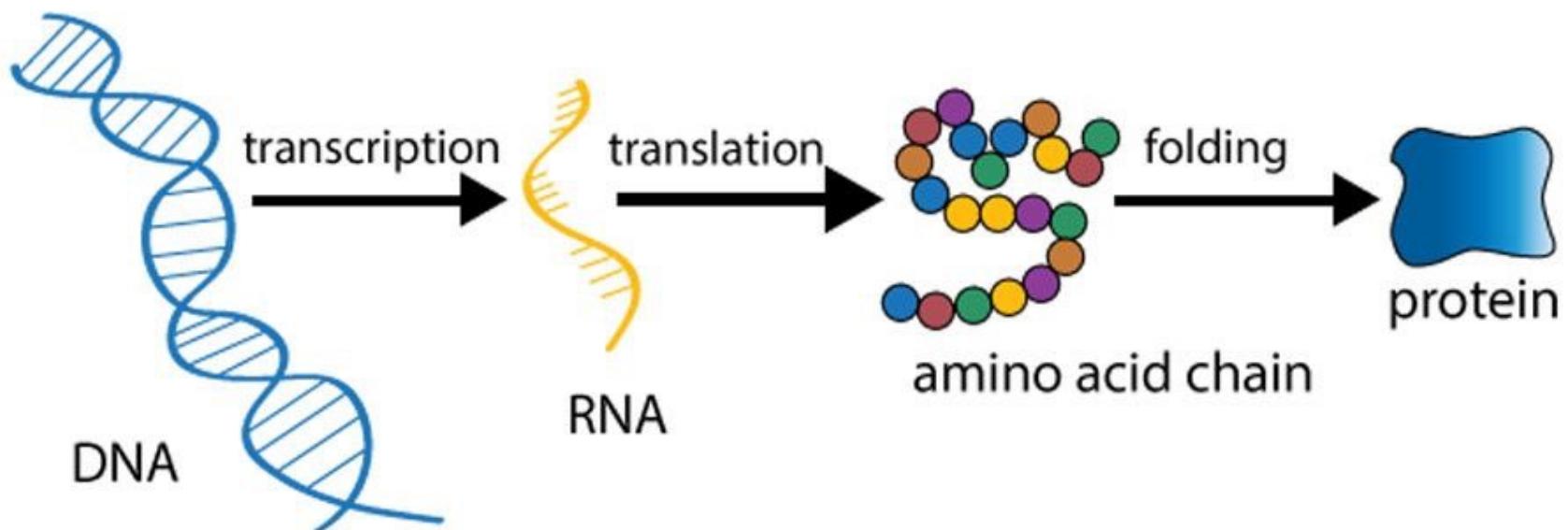
- An overview of genomics and post-genomics era
- Genetic privacy and security
- Ethics and regulations
- Protection techniques

# A Connected Self

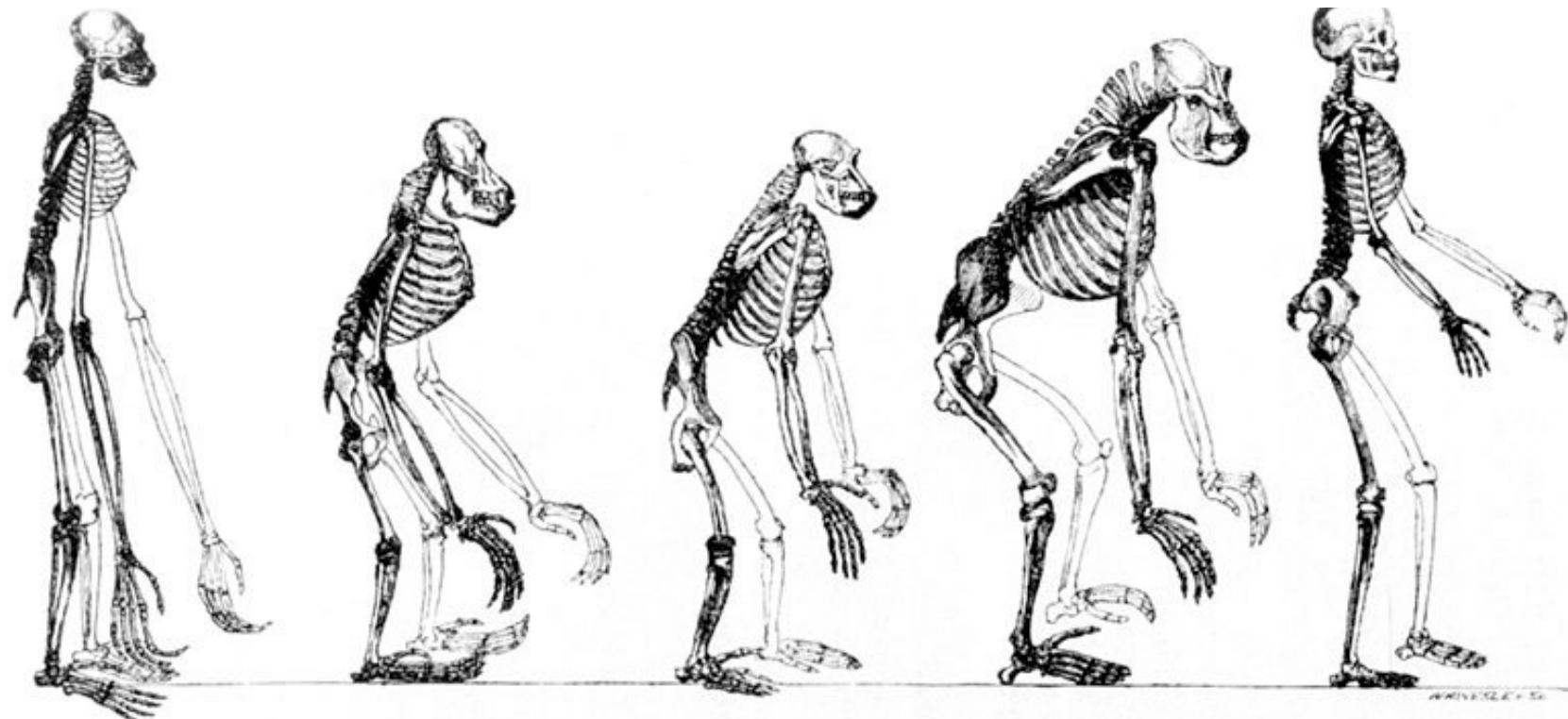


# Genetic and Genomics Data

- Genes, gene products, variants, phenotypes



# Human's Place in Nature



GIBBON.

ORANG.

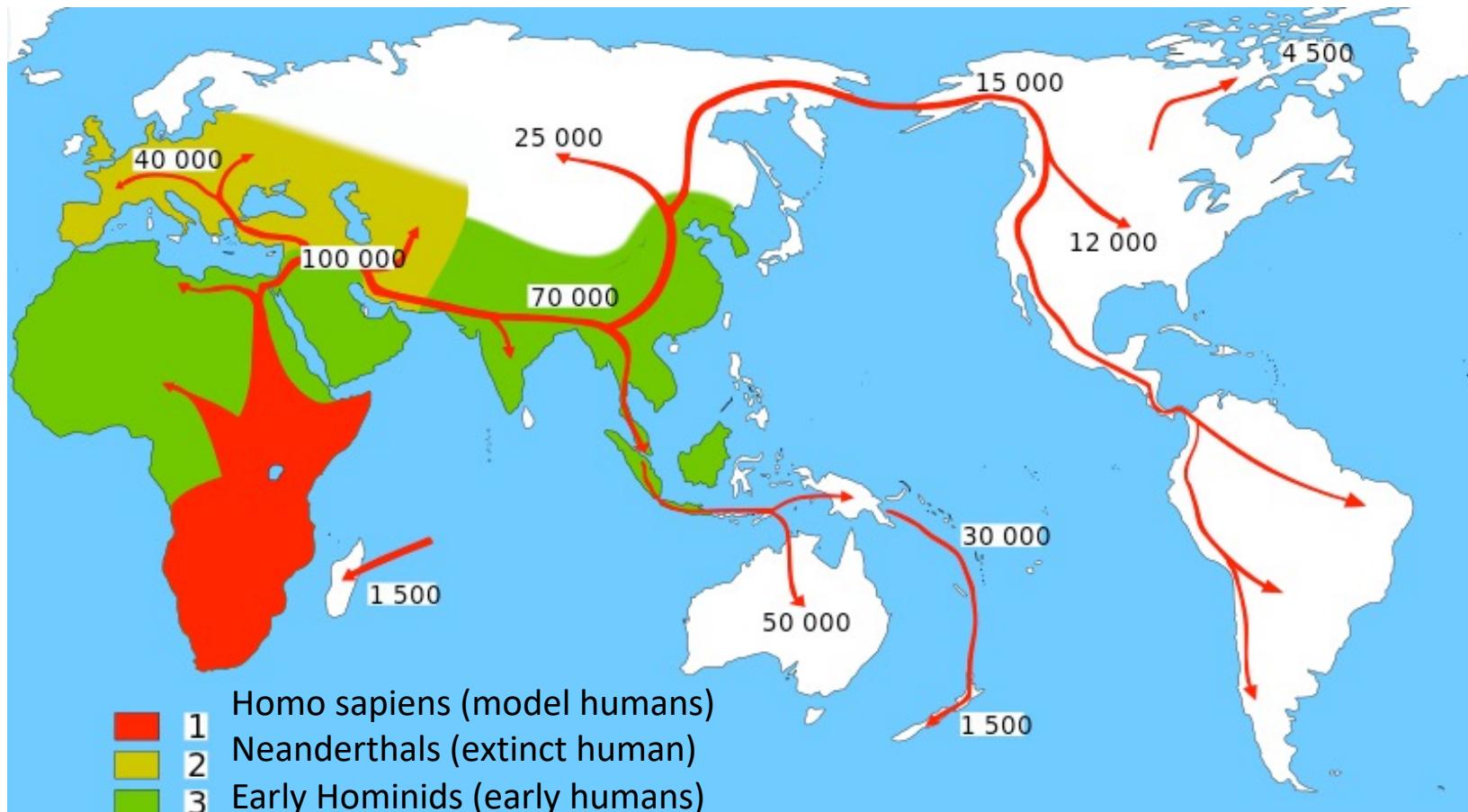
*Skeletons of the  
CHIMPANZEE.*

GORILLA.

MAN.

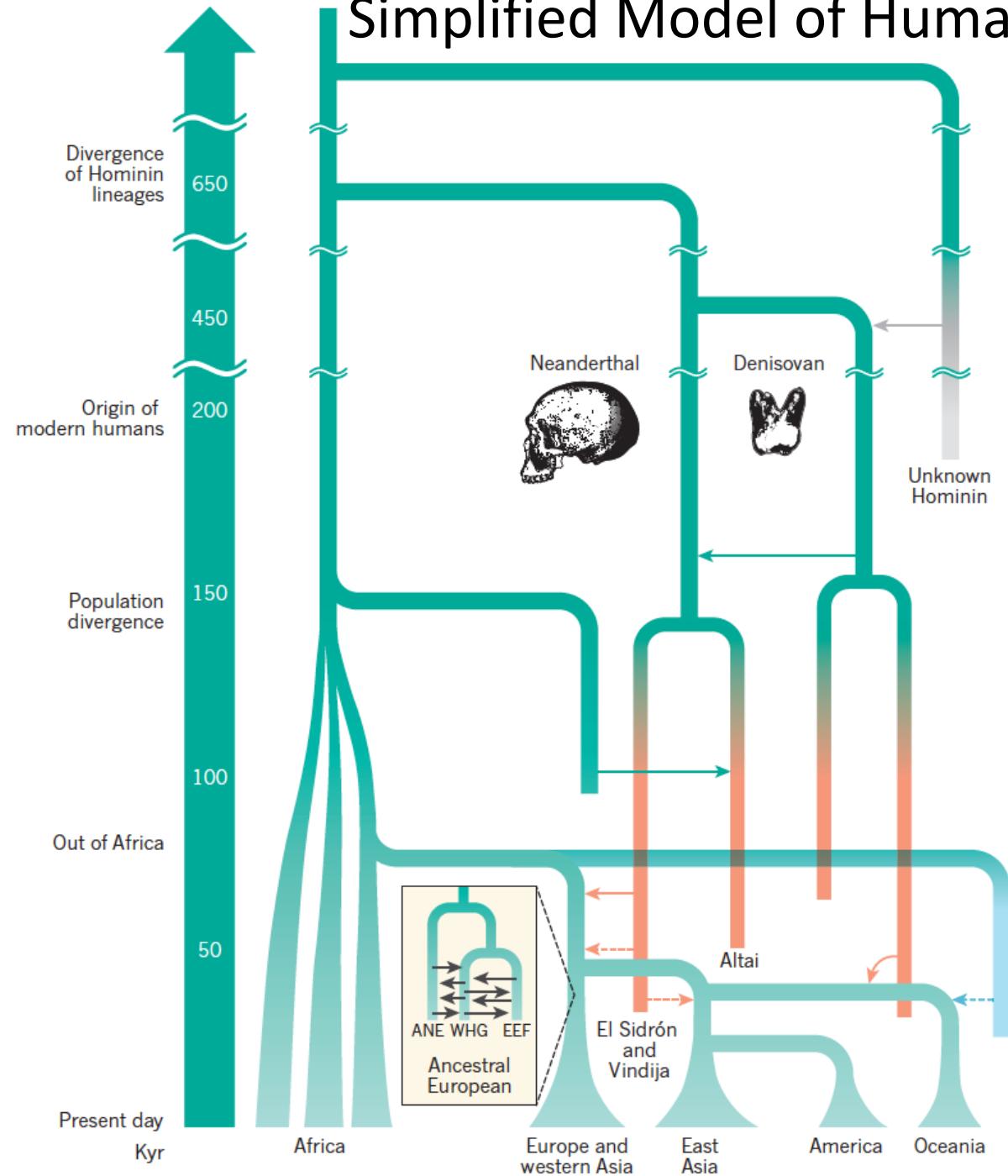
*Photographically reduced from Diagrams of the natural size (except that of the Gibbon, which was twice as large as nature),  
drawn by Mr. Waterhouse Hawkins from specimens in the Museum of the Royal College of Surgeons.*

# Map of early human Migrations “Out-of-Africa”



[http://en.wikipedia.org/wiki/Recent\\_African\\_origin\\_of\\_modern\\_humans](http://en.wikipedia.org/wiki/Recent_African_origin_of_modern_humans)

# Simplified Model of Human Evolutionary History



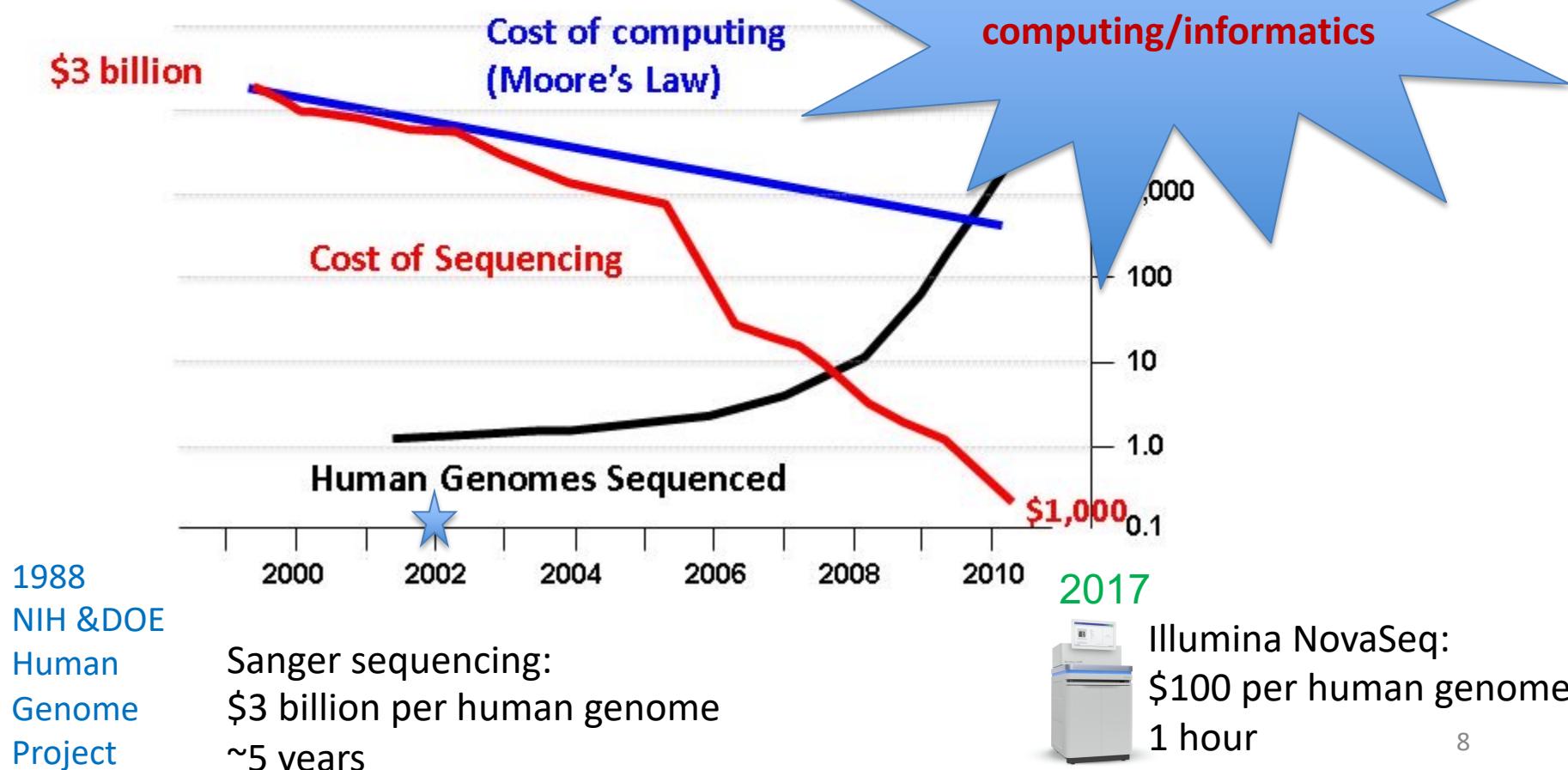
Relationships btw contemporary populations and the approximate times at which they diverged are shown, including admixture events between groups of modern humans and between modern and archaic humans (well established in solid lines and tentative in dashed lines ).

Adapted from

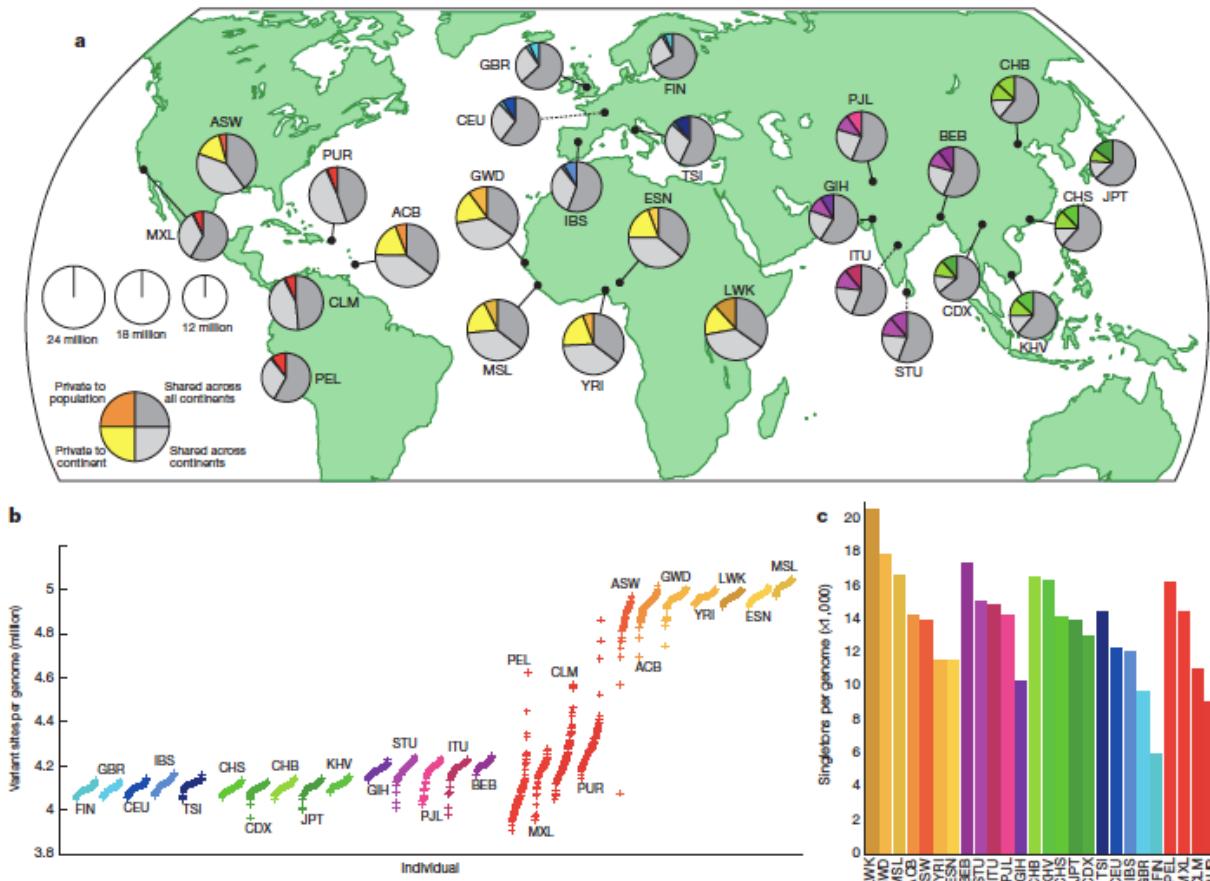
The Economist

# The Sequencing Explosion

Processing this 3-billion-base-pair human genome takes immense computing power



# The 1000 Genomes Project



Whole genome sequencing of 2504 individuals from 26 populations.

- A typical genome differs from the reference human genome at 4.1 ~5.0 million sites (2,100 ~2,500 structural variants).

1000 Genomes Project Consortium. Nature 2010, 2011, 2012, 2015, 2016b.

## SNP-SNV expression quantitative loci mapping analysis:

We identified 54 eQTLs with a lead SV association (denoted SV-eQTL) and 10,100 eQTLs with a lead SNP association (10% FDR). Although SNPs contribute more eQTLs overall, our results suggest SVs have a disproportionate impact on gene expression relative to their number.

# Clinical Sequencing – Federal Initiatives

- International Cancer Genome Consortium (ICCG) and The Cancer Genome Atlas (TCGA) projects chart the genomic changes involved in more than 20 types of cancer (WGS of 5000 individuals, WES of 10,000 individuals).
- Genomes England Project (the 100,000 Genomes Project) 2014, UK 10K Project. -> **UK Biobank**
- Million Genome Project from Obama's Precision Medicine Initiative, 2015. -> **All of Us Project**

# Clinical Sequencing – Private Sections

- Private bio-banks, like 23andMe, have collected spit samples (e.g. “More than 26 million people have taken an at-home ancestry test”).
- Large disease consortia and hospitals/institutions/pharmaceutical/biotech companies conduct whole genome sequencing of clinical samples.

<https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/>

# From Genomics to Metagenomics

- Microbes thrive on us: we provide wonderfully rich and varied homes for our 100 trillion microbial (bacterial and archaeal) partners.
- Human Microbiome Project
  - characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health.
- We are also host to countless viruses. A recent survey reported that human feces contain about a billion RNA viruses per gram, representing 42 viral “species”.
- Viral Metagenomics

National Research Council (US) Committee on Metagenomics:  
Challenges and Functional Applications. 2007

# Outline

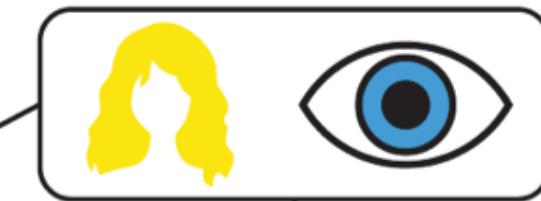
- An overview of genomics and post-genomics era
- **Genetic privacy and security**
- Ethics and Regulations
- Protection techniques

# Why do you care about genetic privacy?

DNA database obtained  
by the adversary



Phenotypic traits of the target(s)  
known to the adversary



Most compatible genotype for  
these phenotypic traits



Infer non-visible phenotypes  
traits from the genotype



e.g., risk of contracting  
Lactose intolerance

# Genetic Privacy

Archive > Volume 493 > Issue 7433 > Editorial > Article

NATURE | EDITORIAL



## Genetic privacy

The ability to identify an individual from their anonymous genome sequence, using a clever algorithm and data from public databases, threatens the principle of subject confidentiality.

17 January 2013

### Routes for breaching and protecting genetic privacy

Yaniv Erlich & Arvind Narayanan

Affiliations | Corresponding author

*Nature Reviews Genetics* 15, 409–421 (2014) | doi:10.1038/nrg3723

Published online 08 May 2014 | Corrected online 17 June 2014

# Genetic Privacy Risks

- Identity tracing attacks
  - e.g. identify a specific individual by their genetic sequence
- Attribute disclosure attacks via DNA
  - e.g. Infer propensity for specific diseases or immediate or distant ancestry
- Completion attacks
  - e.g. complete genetic information from partial data

Erlich Y and Narayanan A, "Routes for breaching and protecting genetic privacy." Nature Reviews Genetics 15.6 (2014): 409.

# Identity Tracing Attacks

- The goal of identity tracing attacks is to uniquely identify an anonymous DNA sample using quasi-identifiers – residual pieces of information that are embedded in the dataset.
- Searching with meta-data
- Identity tracing by genealogical triangulation
- Identity tracing by phenotypic prediction
- Identity tracing by side-channel leaks

# Searching with Meta-data

- Unrestricted demographic information conveys substantial power for identity tracing.
- Health Insurance Portability and Accountability Act (HIPAA) Privacy rule
- Pedigree structures contain rich information, especially when large kinships are available.
- Another vulnerability of pedigrees is combining demographic quasi-identifiers across records to boost identity tracing despite HIPAA protections.

# Searching with Meta-data

- 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}.
- 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes.

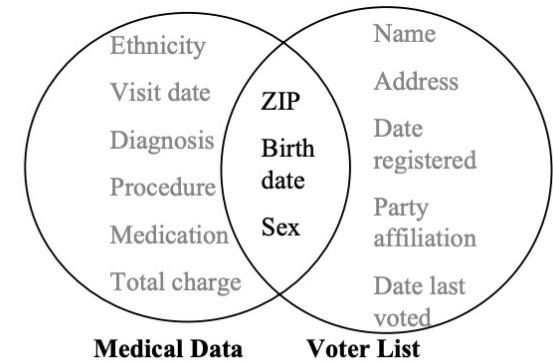


Figure 1 Linking to re-identify data

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.  
Erlich Y and Narayanan A, "Routes for breaching and protecting genetic privacy." Nature Reviews Genetics 15.6 (2014): 409.

# Identity Tracing by Genealogical Triangulation

- Genetic genealogy attracts millions of individuals interested in their ancestry or in discovering distant relatives.
- One potential route of identity tracing is surname inference from Y-chromosome data
- The main limitation of surname inference is that haplotype matching relies on comparing Y chromosome Short Tandem Repeats (Y-STRs).
- An open research question is the utility of non Y chromosome markers for genealogical triangulation.

› [Science](#). 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566.

## Identifying personal genomes by surname inference

# Identity Tracing by Phenotypic Prediction

- Predictions of visible phenotypes from genetic data could serve as quasi-identifiers for identity tracing.
- Twin studies have estimated high heritabilities for various visible traits such as height and facial morphology.
- Age prediction is possible from DNA specimens derived from blood samples.
- But the applicability of these DNA-derived quasi-identifiers for identity tracing has yet to be demonstrated.

# Identity Tracing by Side-channel Leaks

- Side-channel attacks exploit quasi-identifiers that are unintentionally encoded in the database building blocks and structure rather than the actual data that is meant to be public.
- The mechanism to generate database accession numbers can also leak personal information

# Attribute Disclosure Attacks via DNA (ADAD)

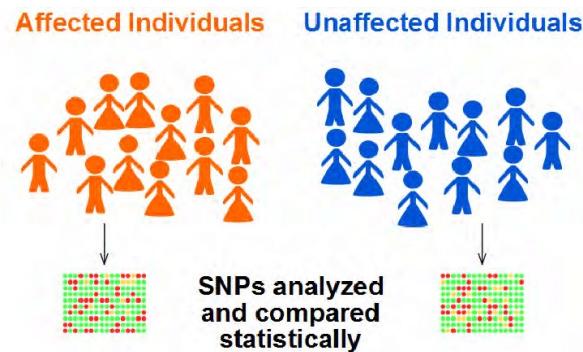
- ADAD attack: The adversary gains access to the DNA sample of the target. He or she uses the identified DNA to search genetic databases with sensitive attributes. A match between the identified DNA and the database links the person and the attribute.
- The simplest scenario
- The summary statistic scenario
- The gene expression scenario

# ADAD: the Simplest Scenario

- The adversary can simply match the genotype data that is associated with the identity of the individual and the genotype data that is associated with the attribute.
- Such an attack requires only a small number of autosomal single nucleotide polymorphisms (SNPs).
- ADAD is a theoretical vulnerability of virtually any individual level DNA-derived omics dataset such as RNA-seq and personal proteomics.
- Genome-wide association studies (GWAS) are highly vulnerable to ADAD.

# ADAD: the Summary Statistic Scenario

- With the target genotypes in the case group, the allele frequencies will be positively biased towards the target genotypes compared to the allele frequencies of the general population.
- The actual risk of ADAD has been the subject of intense debate.



# ADAD: the Gene Expression Scenario

- NIH's Gene Expression Omnibus (GEO) publicly hold hundreds of thousands of gene expression profiles.
- ADAD technique: The method starts with a training step that employs a standard expression quantitative trait loci (eQTL) analysis with a reference dataset. Next, the algorithm scans the public expression profiles. Last, the algorithm matches the target's genotype with the inferred allelic distributions of each expression profile and tests the hypothesis that the match is random.
- This ADAD technique has the potential for relatively high accuracy in ideal conditions.

> [Nat Methods](#). 2016 Mar;13(3):251-6. doi: 10.1038/nmeth.3746. Epub 2016 Feb 1.

Quantification of private information leakage from phenotype-genotype data: linking attacks

Arif Harmanci [1](#) [2](#), Mark Gerstein [1](#) [2](#) [3](#)

# Completion Attacks

- Genotype imputation: Jim Watson's predisposition for Alzheimer's disease from the ApoE locus despite masking of this gene
- In the ***basic*** setting, the adversary obtains access to a single genetic dataset of a known individual. He then exploits this information to estimate genetic predispositions for relatives whose genetic information is inaccessible.
- In the ***advanced*** setting, the adversary has access to the genealogical and genetic information of multiple relatives of the target. The algorithm finds relatives of the target that donated their DNA to the reference panel and that reside on a unique genealogical path that includes the target.

# Homer's Attack

- This paper demonstrated the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture.
- Homer's attack motivated the NIH to move the genotype and phenotype data from public domain to controlled access through dbGaP.

Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008.

# Genomic Data

# Privacy Breaches

# Privacy Protection

## Private/Controlled

dbGaP  
EBI  
Biobanks  
Hospitals

Genealogy databases  
Commercial genetic databases

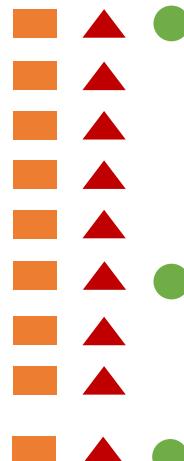
## Public

HapMap  
1000GP  
HGDP  
SGDP  
openSNP  
HGP

GWAS catalog

Identity attack   Trait attack   Completion attack

## Genotypes, Statistics and Metadata



- Wheeler 2008  
Homer 2008  
Wang 2009  
Shringapyre 2015  
Wang 2013, 2015, 2016  
Nyholt 2009  
Humbert 2015, 2017  
Ney 2019  
Ayoz 2021

- Goodrich 2009  
Gymrek 2013  
Erlich 2018  
Edge 2019

## RNA and Statistics



- Harmanci 2016

## Ethics/Regulations/Access Control

Ethics education

Health Insurance Portability and Accountability Act (HIPAA) Breaches (1996)

## Differential Privacy

GWAS logistic regression

Advanced statistical and ML models

## Cryptography

Homographic encryption

Secure multiparty computation

## Trusted Executive Environment

Intel SGX

# The Golden State Cold Case

- The DNA profile of the Golden State Killer was uploaded to GEDmatch, an open-source platform frequently used by members of the public to trace their heritage.
- The test result was first sent to FamilyTreeDNA, which created a DNA profile and allowed law enforcement to set up a fake account to search for matching customers.
- When that produced only distant leads, a civilian geneticist working with investigators uploaded the forensic profile to MyHeritage.
- It was the MyHeritage search that identified the close relative who helped break the case.

Los Angeles Times

SUBSCRIBE



CALIFORNIA

The untold story of how the Golden State Killer was found: A covert operation and private DNA

# Ethics and Privacy

NEWS CAREERS COMMENTARY JOURNALS ▾

Science

## We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans

Researchers call for limiting how ancestry databases can be used to protect privacy

---

11 OCT 2018 • BY JOCELYN KAISER

# Outline

- An overview of genomics and post-genomics era
- Genetic privacy and security
- Ethics and regulations
- Protection techniques

# Ethics of genetic privacy and current regulations

- Standards for Privacy of Individually Identifiable Health Information in the Health Insurance Portability and Accountability Act of 1996 (HIPAA).
- The Privacy Rule was established to address the use and disclosure of individuals' health information by covered entities, and provides standards for individual privacy rights to understand and control the use of their health information.
- Many large human genome projects provide ethics education

# Ethics and HIPAA Review

- Key to advancing genetics diagnosis research
- Private personal health information can be protected
- Discrimination/Bias based on released heath information can be eliminated (minimized)
- HIPAA Privacy Rule: All federal grants with human subjects involved should be protected by HIPAA

<https://www.hhs.gov/hipaa/index.html>

# Introduction to HIPAA

- The Standards for Privacy of Individually Identifiable Health Information (“Privacy Rule”) establishes, for the first time, a set of national standards for the protection of certain health information.
- The U.S. Department of Health and Human Services (“HHS”) issued the Privacy Rule to implement the requirement of the Health Insurance Portability and Accountability Act of 1996 (“HIPAA”).
- Within HHS, the Office for Civil Rights (“OCR”) has responsibility for implementing and enforcing the Privacy Rule with respect to voluntary compliance activities and civil money penalties.

# Information Protected by HIPAA

- Protected Health Information
  - The Privacy Rule protects all “**individually identifiable health information**”
  - Only covers patient information kept by “**covered entities**”, i.e., health providers, insurers and data clearinghouses, as well as their business partners
- De-Identified Health Information
  - There are no restrictions on the use or disclosure of de-identified health information.

# Information NOT Protected by HIPPA

- De-identified health information
- Medical information not originated from entities not covered by HIPAA (e.g. 23andMe, genetic screening companies)
- Consumer-generated health information (home paternity tests, fitness trackers, health apps, social media)
- Personal information (ethic, identity, etc.) inferred from de-identified health information
- Meta data such as age, geographical regions, races of participants

# HIPAA Breaches

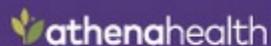
More than 41 million people have had their protected health information compromised in HIPAA privacy and security breaches.



TOPICS

SIGN UP

MAIN MENU



We connect patients and providers in the cloud. **We connect care.**

Expand +

Privacy & Security

## HIPAA breaches: The list keeps growing

Our searchable tally of HIPAA breaches since 2009 shows an industry still unprepared to keep data safe

<https://www.hhs.gov/hipaa/index.html>

# NIH Policy and Ethics Issues

- Coverage and Reimbursement of Genetic Tests
- Genetic Discrimination
- Informed Consent for Genomic Research
- Intellectual Property
- Privacy in Genomics
- Regulation of Genetic Tests

<https://www.genome.gov/27527631/policy-ethics-issues/>

# Coverage and Reimbursement of Genetic Tests

- Genomic medicine has the capacity to revolutionize clinical practice.
- One challenge insurers face is the difficulty of deciding when to reimburse for genetic tests that health care providers have offered their patients.
- Payers are having trouble keeping up with the volume of new genetic and next-generation sequencing tests that are coming onto the market.

<https://www.genome.gov/19016729/coverage-and-reimbursement-of-genetic-tests/>

# HIPAA Safe Harbor Rule

- Dissemination of demographic identifiers has been the subject of tight regulation in the US health care system.
- The maximal resolution of any date field, such as hospital admission dates, is in years.
- The maximal resolution of a geographical subdivision is the first three digits of a zip code (for zip code areas with populations of >20,000).

## Genetic Information Nondiscrimination Act of 2008



**Long title**

An act to prohibit discrimination on the basis of genetic information with respect to health insurance and employment.

**Acronyms  
(colloquial)**

GINA

**Enacted by  
Effective**

the 110th United States Congress  
May 21, 2008

<http://www.ginahelp.org/GINAhelp.pdf>

# Genetic Discrimination

- Many Americans fear that participating in research or undergoing genetic testing will lead to being discriminated against based on their genetics.
- The Genetic Information Nondiscrimination Act (GINA) was passed into law, prohibiting discrimination by employers and health insurers.
- There are also other legal protections against genetic discrimination by employers, health insurers, and others.

# Informed Consent for Genomic Research

- Advances in genomic technology and analytical tools are enabling discoveries that enhance our understanding of the impact of genomic variants on health and disease.
- Informed consent shows respect for personal autonomy and is an important ethical requirement in research. (HIPAA Privacy Rule)
- Informed consent involves two fundamental components: a dialogue or process, and a form.

<https://www.genome.gov/27026588/informed-consent-for-genomics-research/>

# Intellectual Property

- In a landmark decision in June 2013, the Supreme Court determined that DNA in its natural form cannot be patented.
- The National Human Genome Research Institute (NHGRI) considers research "legal issues regarding patents" as part of the center's research into the ethical, social, and legal implications of human genome research.
- The Courts and Gene Patents

<https://www.genome.gov/19016590/intellectual-property/>

# Privacy in Genomics

- Each person's DNA sequence includes health and other information about them and their families.
- Usage and privacy need to be balanced.
- Genetic Information Nondiscrimination Act (GINA)
- The HIPAA Privacy Rule
- Certificates of Confidentiality
- The Freedom of Information Act (FOIA) gives any person the right to request access to records of the Executive Branch of the United States Government
- NIH Genomic Data Sharing Policy

# Outline

- An overview of genomics and post-genomics era
- Genetic privacy and security
- Ethics and regulations
- Protection techniques

# Techniques for Privacy Protection

- Access control
- Differential privacy (DP)
- Cryptographic solutions
  - Homomorphic encryption (HE)
  - Secure multiparty computation (MPC)

# Access Control

- Allows users to download data only after approval (e.g. dbGaP)
- A trust-but-verify approach: where users cannot download the data without restriction but may execute certain types of queries, which are recorded and audited by the system (e.g., GA4GH)
- Allowing the original participants to grant access to their data instead of delegating this responsibility to a data access committee.



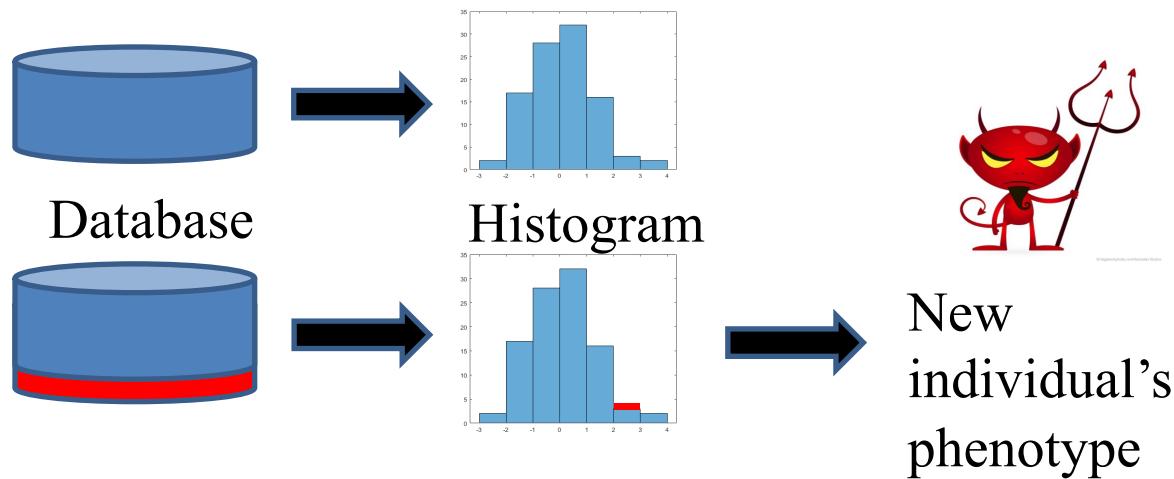
ga4gh.org



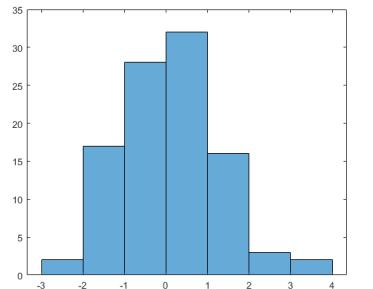
**Global Alliance**  
for Genomics & Health  
Collaborate. Innovate. Accelerate.

# Differential Privacy (DP)

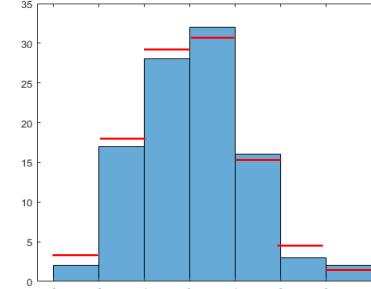
- Differential privacy is a system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.



# Differential Guarantee

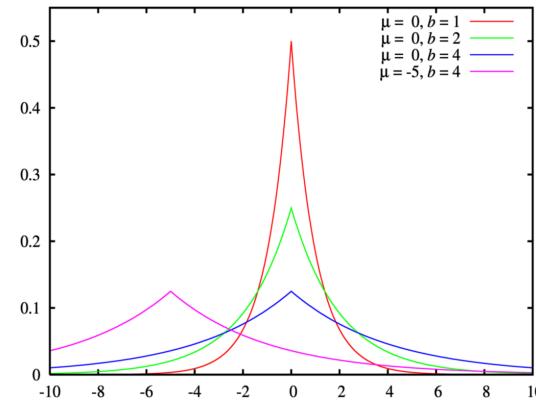


+ noise



noise  $\sim$  Laplacian distribution

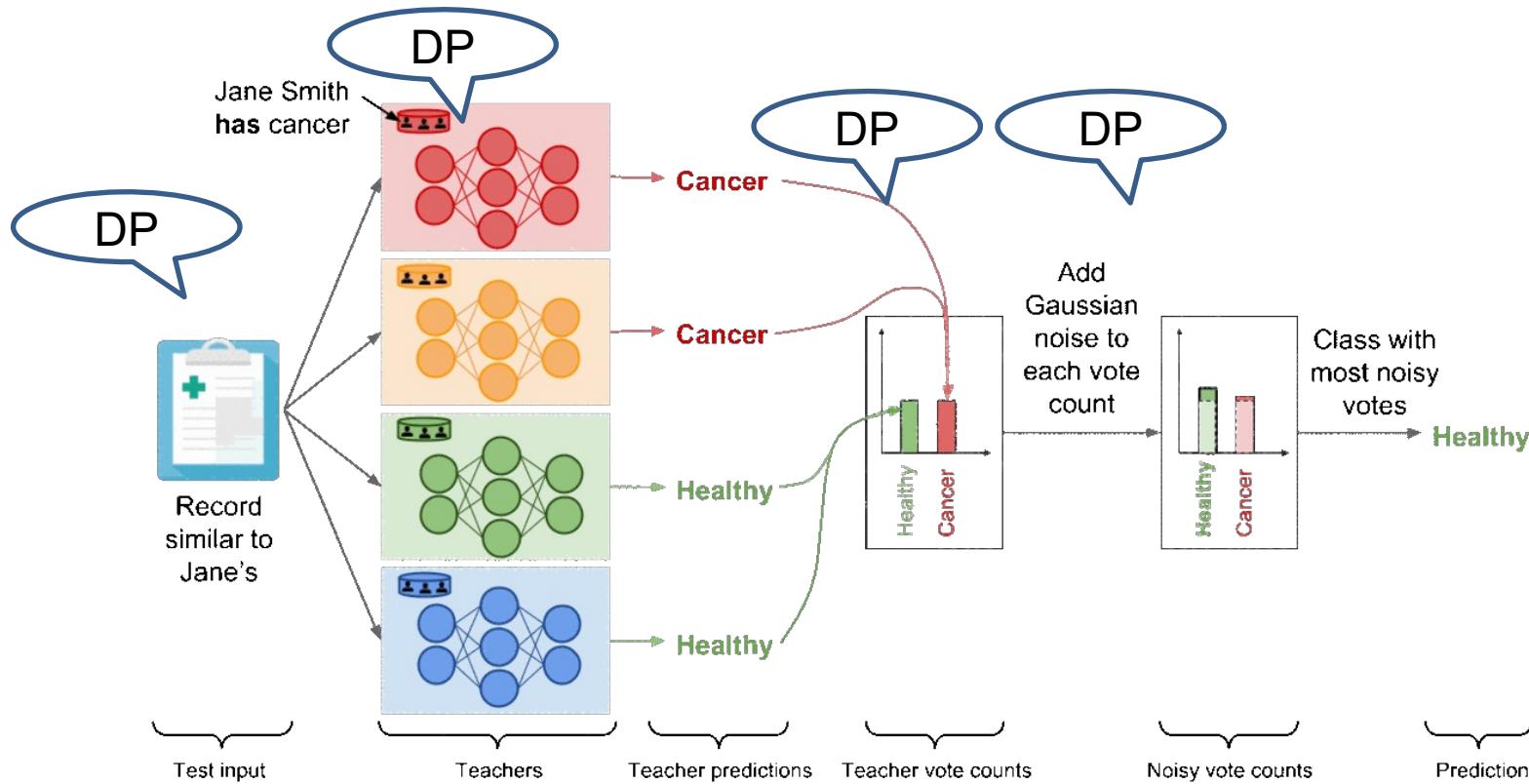
$$P(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$



Dwork C, and Aaron R. "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014): 211-407.

Task C. "Privacy-preserving social network analysis", Purdue University.

# Differential Privacy in Machine Learning



# Homomorphic Encryption (HE)

- Homomorphic encryption is a form of encryption that allows computations to be carried out on ciphertext, thus generating an encrypted result which, when decrypted, matches the result of operations performed on the plaintext.

$$f(x @ y) = f(x) @ f(y)$$

where  $@$  can be any operator.

- Let's define our notation for message, ciphertext, encryption, and decryption:

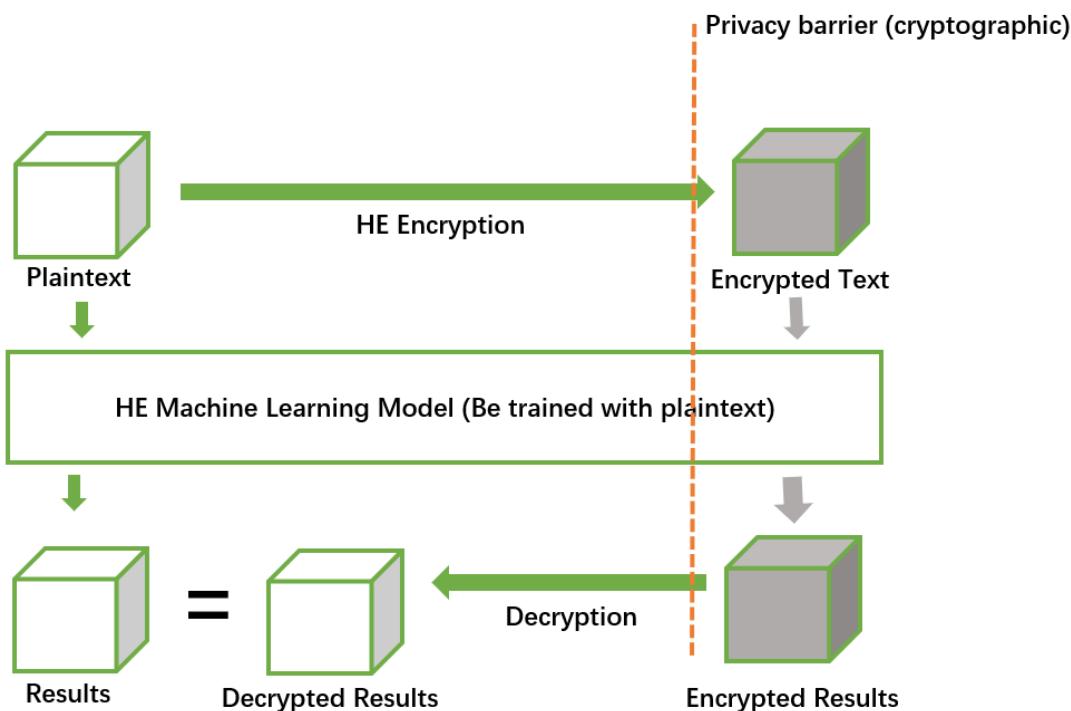
$$\text{Encryption} : E(m) = c$$

$$\text{Decryption} : D(c) = m$$

- Assuming homomorphism, we then get:

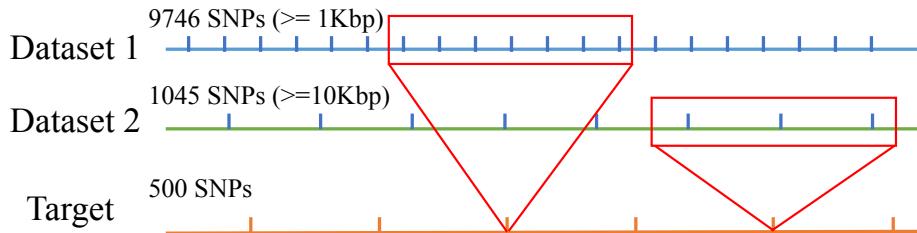
$$E(m_1) + E(m_2) = E(m_1 + m_2) \equiv D(E(m_1 + m_2)) = m_1 + m_2$$

# Secure and Privacy-preserving ML via HE

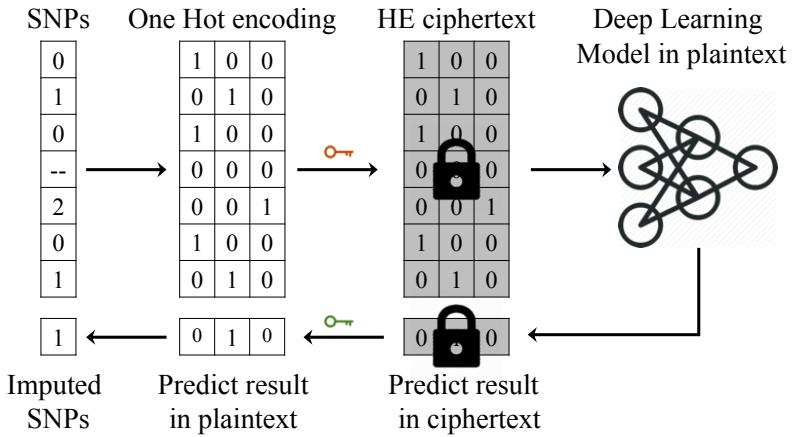


1. Machine Learning models will be trained using plaintext.
2. The trained plain models can be used on homomorphically encrypted data, so that it can be hosted on untrusted servers.
3. After decrypting, the results are same as results of non-encrypted data

# iDASH 2019: HE-Imputation



An illustration of genotype imputation as a classification problem. Missing SNPs are imputed by using their adjacent SNPs.



SNPs representation and imputation workflow.

IDASH PRIVACY & SECURITY WORKSHOP 2019 -  
secure genome analysis competition

\*NHGRI R13HG009072

# Secure Multiparty Computation (SMC)

## Yao's millionaires' Problem



Alice



Bob

This problem discusses two millionaires, Alice and Bob, who are interested in knowing which of them is richer without revealing their actual wealth.

Solution:  
The trusted third person



# Secure Multiparty Computation (SMC)

## Private auction

- Many parties wish to execute a private auction
- The highest bid wins
- Only the highest bid (and bidder) is revealed



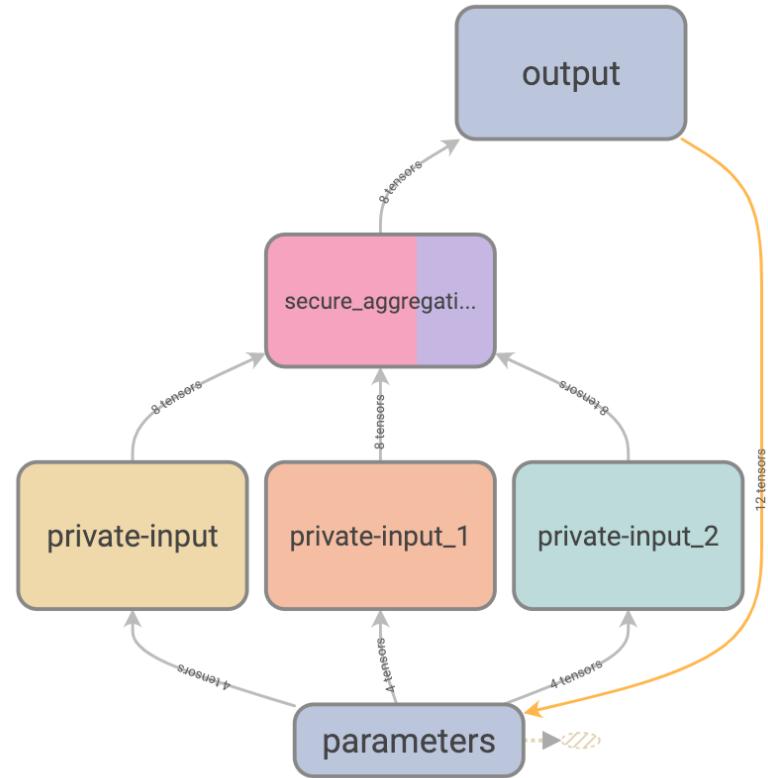
Solution:  
a trusted auctioneer



# Secure Multiparty Computation

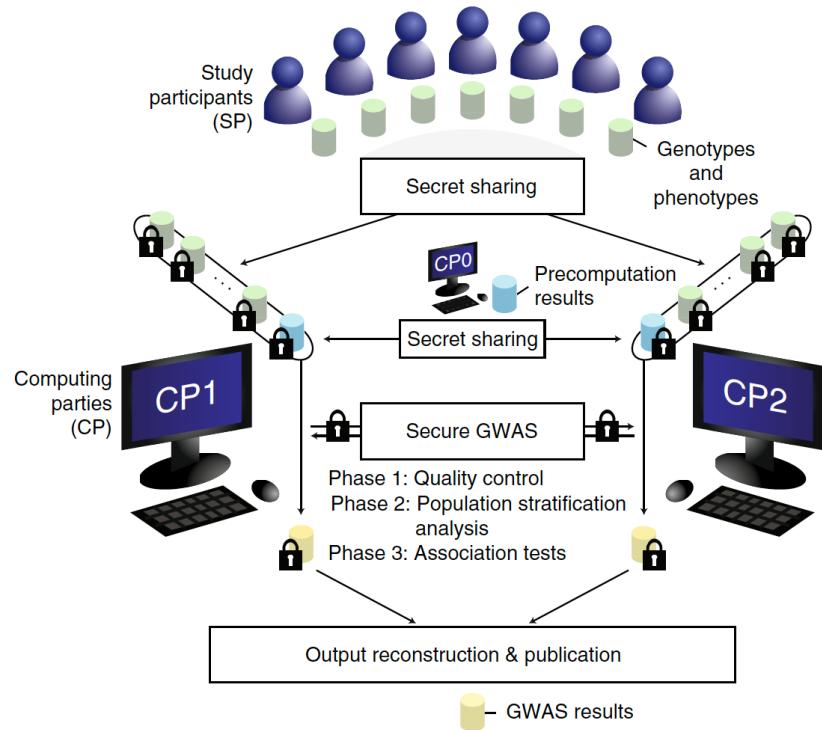
Secure multiparty computation allows us to perform analysis on private data without compromising it.

- Parties  $P_1, \dots, P_n$
- Party  $P_i$  has private input  $x_i$
- The parties wish to jointly compute a (known/unknown) function  $y = f(x_1, \dots, x_n)$
- The computation must preserve certain security properties, even if some of the parties collude and maliciously attack the protocol.



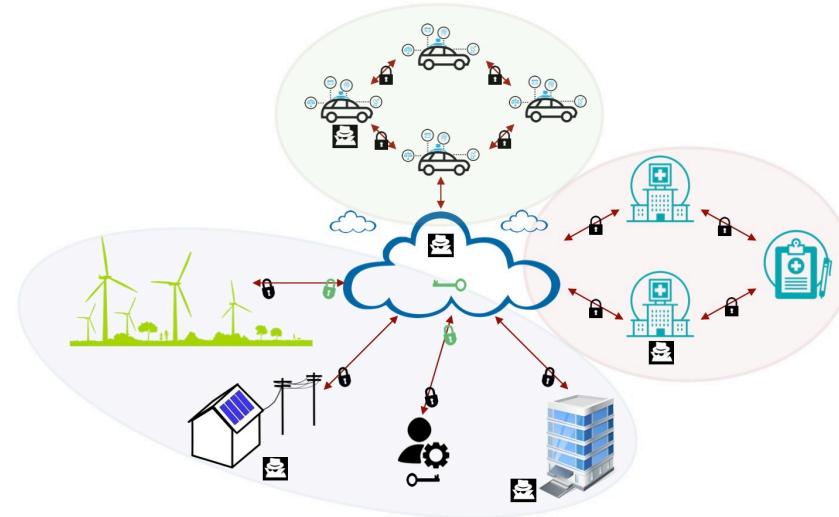
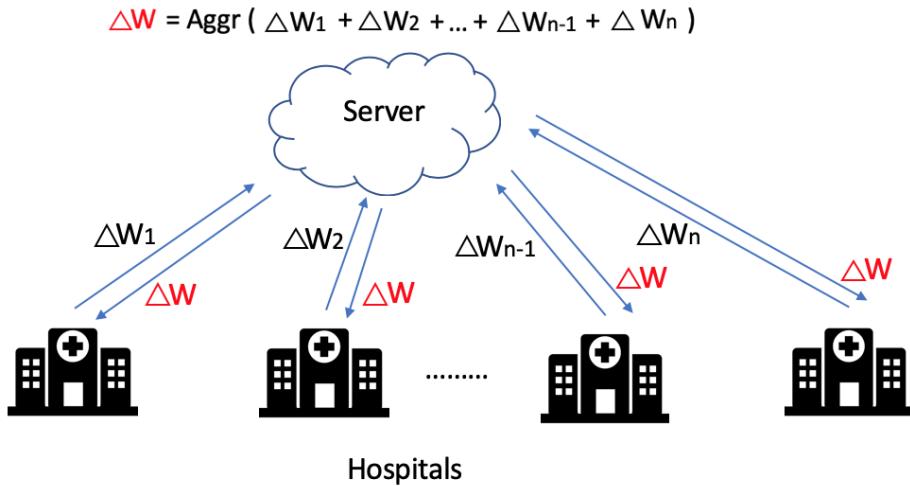
# Secure GWAS Using SMC

- Study participants (private individuals or institutes) secretly share their genotypes and phenotypes with computing parties (research groups or government agencies), denoted CP1 and CP2.
- CP1 and CP2 jointly carry out the secure genome-wide association study (GWAS) protocol to obtain association statistics without revealing the underlying data to any party involved.
- An auxiliary computing party (CP0) performs input-independent precomputation to greatly speed up the main computation



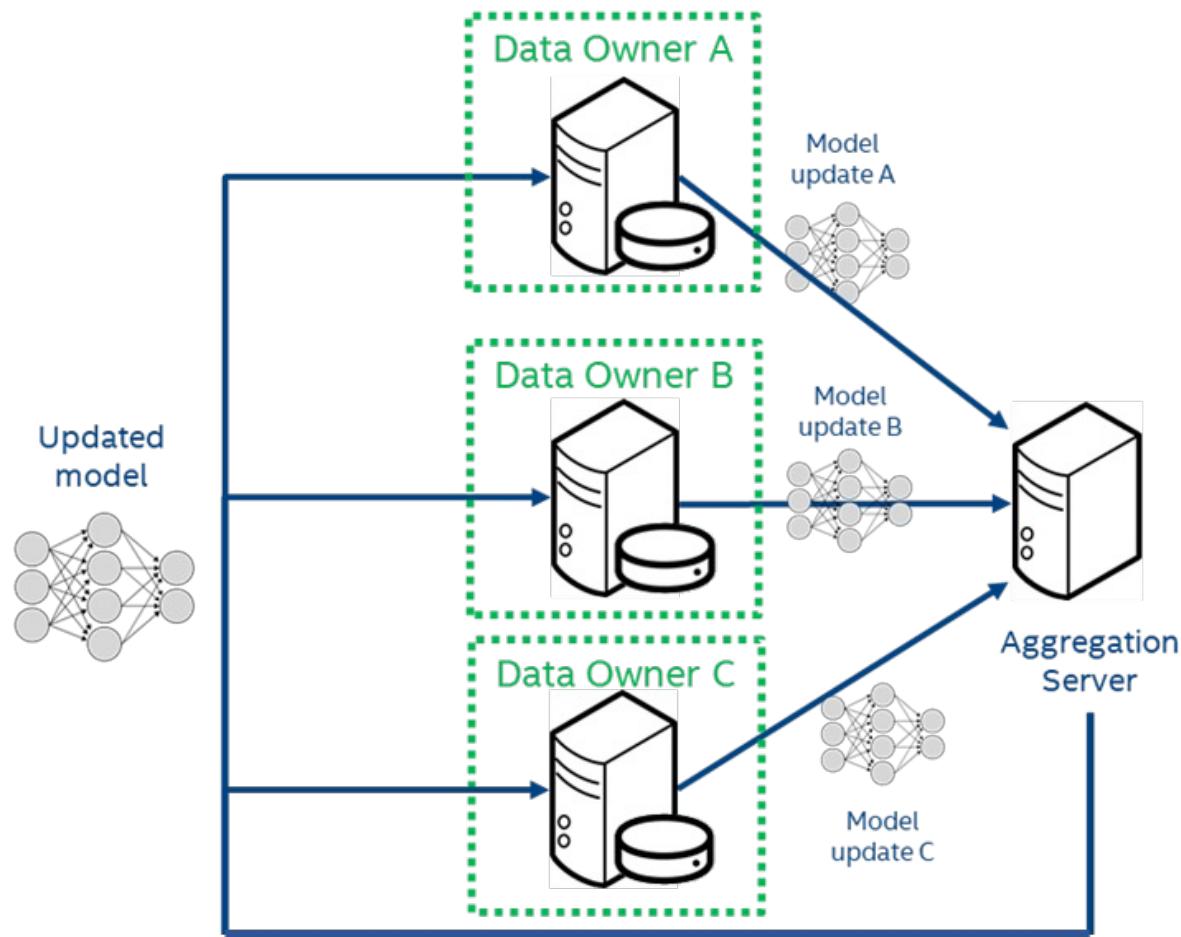
Cho H, Wu DJ, Berger B, "Secure genome-wide association analysis using multiparty computation." Nature Biotechnol. 2018 Jul;36(6):547-551.

# Federated Learning

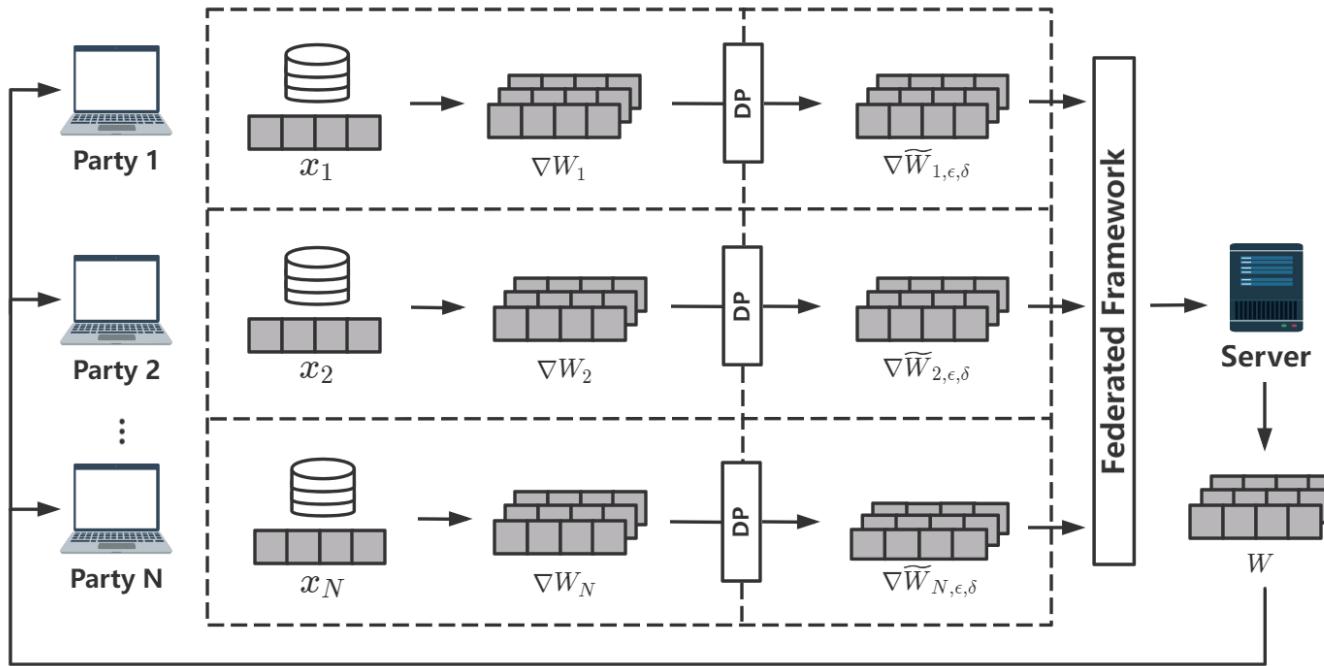


**Federated learning (FL)** is a learning paradigm seeking to address the problem of data governance and privacy by training algorithms collaboratively without exchanging the data itself.

# Hybrid Infrastructure for Privacy Preserving Distributed Machine Learning



# FedDP: Differentially Private FL for Disease Prediction



**An Overall Framework of FedDP.** Local data  $x_i$  distributed at each party is independent from each other. First, local gradients  $\nabla W_i$  are generated in parallel by each local optimizer deployed at each party and trained on each local dataset. Then, the server gathers all gradients  $\nabla W_{i,\epsilon,\delta}$  from multiple parties after  $(\epsilon, \delta)$  DP was applied to original gradients  $\nabla W_i$ . Finally, global weights are computed and broadcast to each party at the end of each round by aggregating all  $\nabla \widetilde{W}_{i,\epsilon,\delta}$ .

# Summary of Protection Techniques

- A hybrid system may need to integrate multiple techniques
- Tradeoff between utility and privacy/security
- These methods are computationally intensive, and hard to scale up to large datasets
- Confidential Computing with Trusted Executive Environment (e.g. Intel's SGX)

# Conclusion

- Security/privacy of genomic data is a growing concern
- Existing regulations and techniques are not sufficient for protecting genetic privacy
- New regulations, guidelines, and techniques are to be developed to realize the full potential of genomic medicine and precision health
- Research and education from multiple disciplines is in great need to advance precision health and open science