PROGRESS OF HOMOMORPHIC ENCRYPTION FOR PROTECTING GENOMIC DATA PRIVACY AND SECURITY IN THE PAST 4 YEARS IDASH COMPETITION

SHUANG WANG, PHD

DEPARTMENT OF BIOMEDICAL INFORMATICS

UNIVERSITY OF CALIFORNIA SAN DIEGO

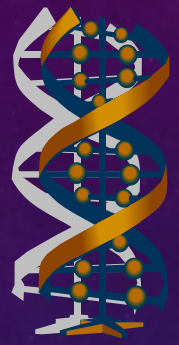
SUPPORTED BY

NHGRI R00HG008175, R13HG009072

NIBIB U01EB023685

Joint work with Dr. Xiaoqian Jiang (UCSD), Dr. Lucila, Ohno-Machado (UCSD), Xiaofeng Wang (IU), Haixu Tang (IU)

HUMAN GENOME PRIVACY



- Genome data have been widely used in biomedical research
- But genomic data are also highly sensitive
  - Diseases association: predisposition to Diabetes, Cancer…
  - Re-identification: name…
  - Information disclosure of blood relatives
  - **A great fear of unknown**

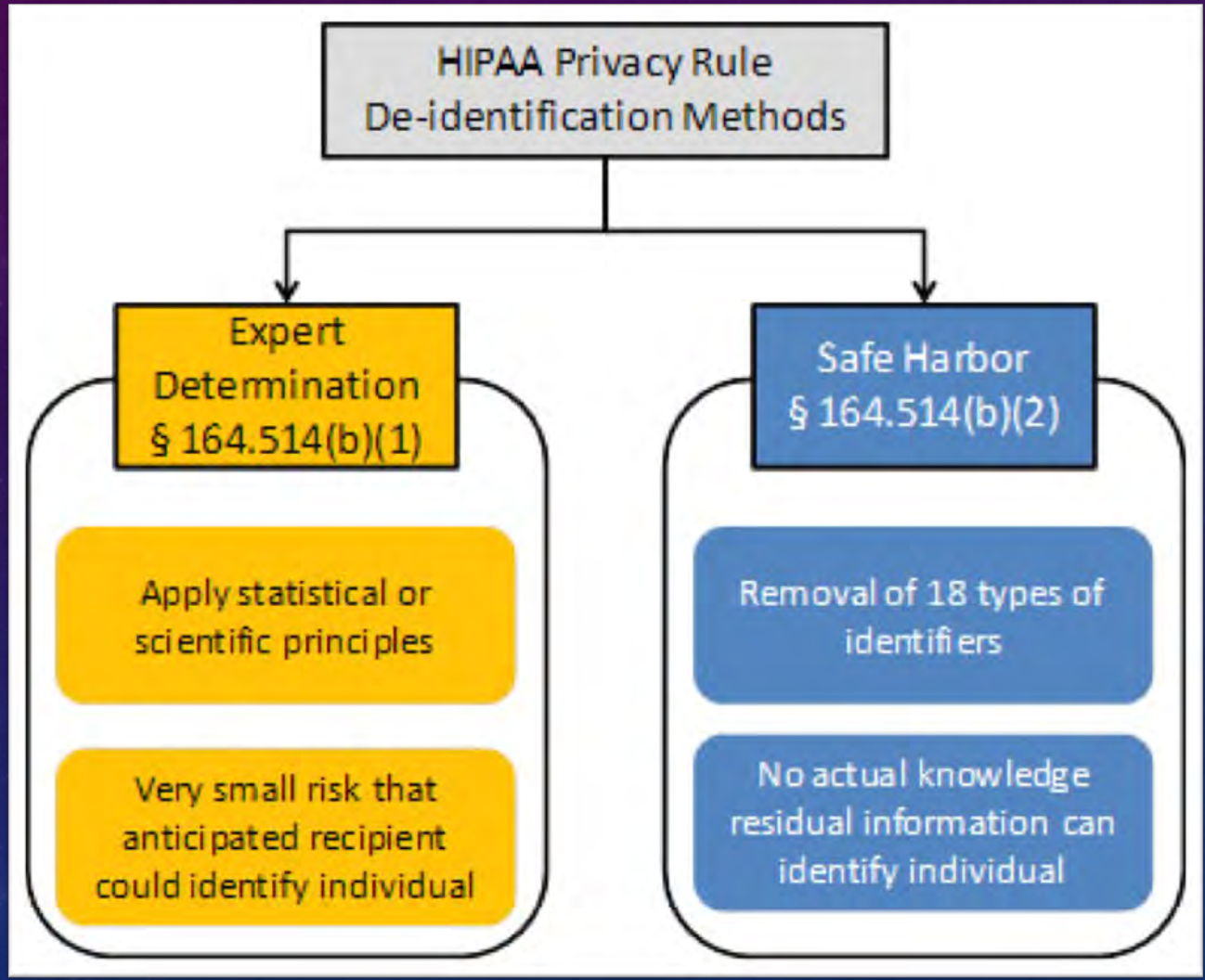
Table 1: A summary of some existing privacy risks to biomedical and genomic data

Author	Year	Summary
Sweeney [6]	2000	*Identifying 87% of US citizens with the combination of 'ZIP code, gender, date of birth'*
Gottlib [7]	2001	*Finding employees who are susceptible to genetic diseases depending on genomic data*
Lin et al. [8]	2004	*Identifying a person by as few as 75 independent SNPs*
Homer et al. [9]	2008	*Detecting if an individual is present in a DNA mixture within a case group*
Sankararaman et al. [10]	2009	
Wang et al. [11]	2009	*Re-identifying individuals and reconstructing allele frequencies from research papers*
Gymrek et al. [12]	2013	*Identifying surnames by profiling short tandem repeats on the Y-chromosome*
Claes et al. [13]	2014	*Reconstructing a 3D face from human genomic data*
Shringarpure et al. [14]	2015	*Identifying participants from using Beacon services with limited number of queries*
Harmanci et al. [15]	2016	*Linking phenotype and genotype data to reveal private information*
Lippert et al. [16]	2017	*Identification of individuals by trait prediction using whole-genome sequencing data*

President Obama Speaks on the Precision Medicine Initiative

HIPAA REGULATES MEDICAL DATA SHARING

HIPAA: Health Insurance Portability and Accountability Act



Nass SJ, Levit LA, Gostin LO. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. The National Academies Press; 2009.

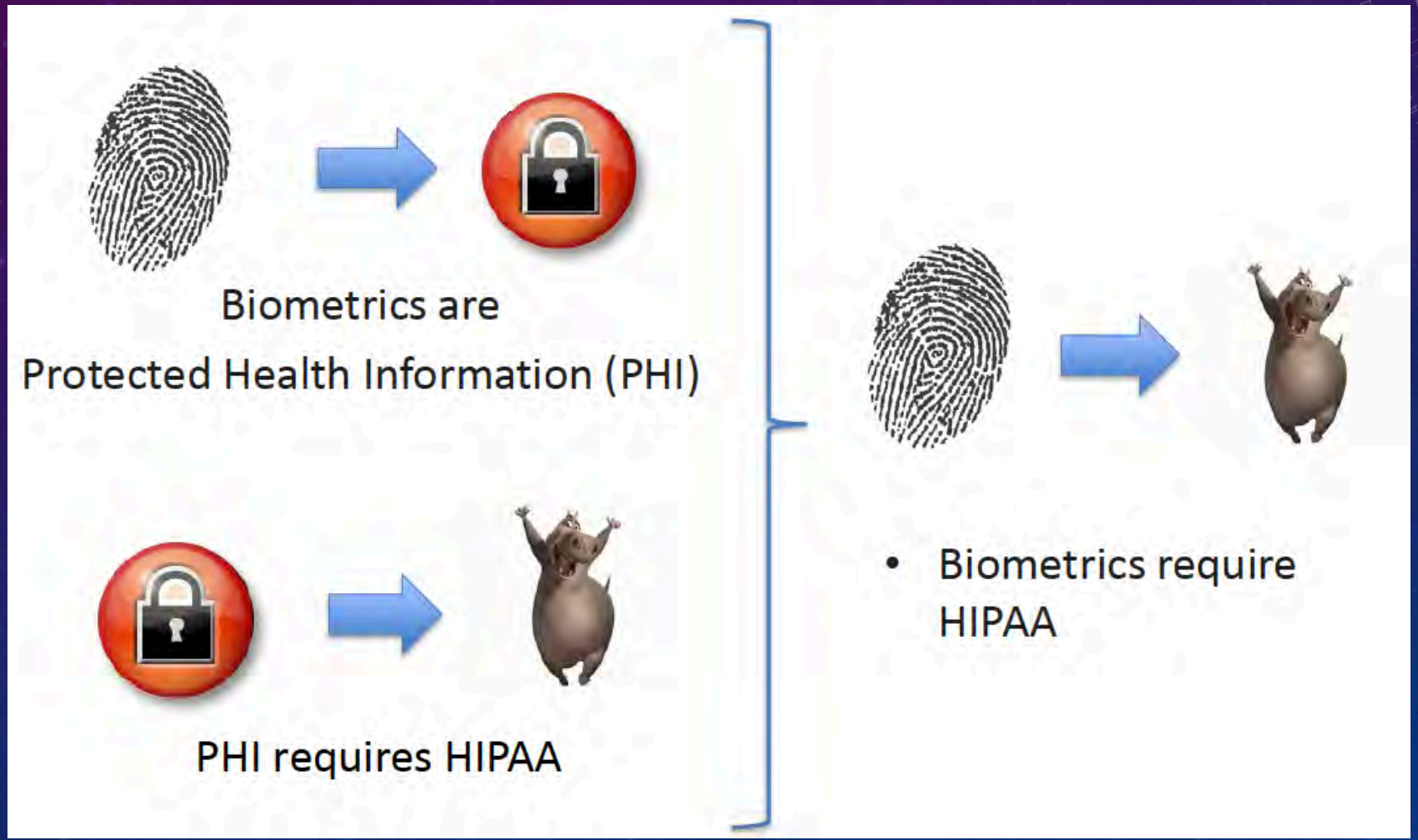
Expert Determination SAFE HARBOR

"A person with **appropriate knowledge** of and **experience** with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable"

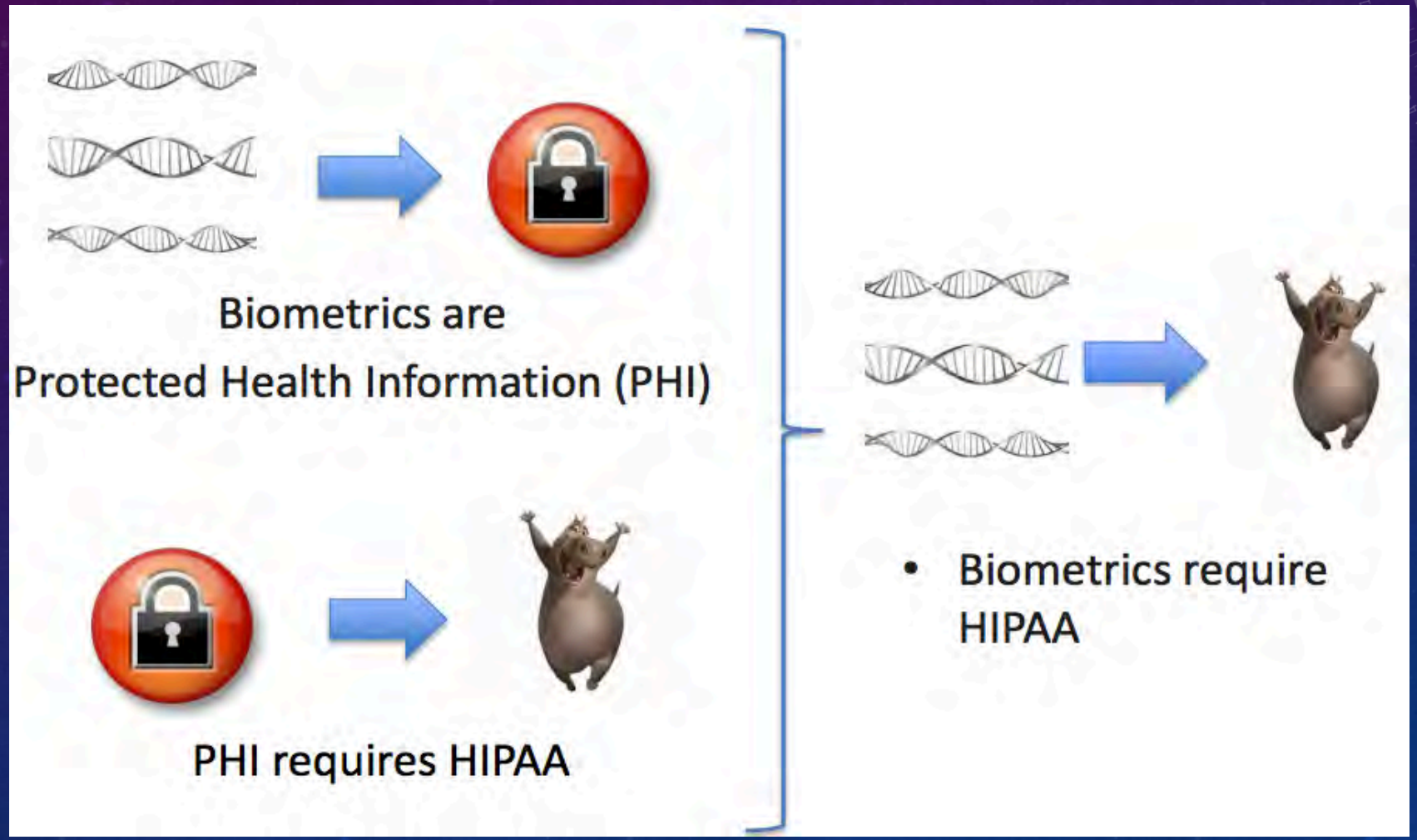
This method is seldom used in practice

<table>
<tr><td colspan="2">(A) Names</td></tr>
<tr><td colspan="2">(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
(1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
(2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000</td></tr>
<tr><td colspan="2">(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older</td></tr>
<tr><td>(D) Telephone numbers</td><td>(L) Vehicle identifiers and serial numbers, including license plate numbers</td></tr>
<tr><td>(E) Fax numbers</td><td>(M) Device identifiers and serial numbers</td></tr>
<tr><td>(F) Email addresses</td><td>(N) Web Universal Resource Locators (URLs)</td></tr>
<tr><td>(G) Social security numbers</td><td>(O) Internet Protocol (IP) addresses</td></tr>
<tr><td>(H) Medical record numbers</td><td>(P) Biometric identifiers, including finger and voice prints</td></tr>
<tr><td>(I) Health plan beneficiary numbers</td><td>(Q) Full-face photographs and any comparable images</td></tr>
<tr><td>(J) Account numbers</td><td rowspan="2">(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and</td></tr>
<tr><td>(K) Certificate/license numbers</td></tr>
<tr><td colspan="2">(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.</td></tr>
</table>

PROBLEMS WITH THE SAFE HARBOR METHOD



PROBLEMS WITH THE SAFE HARBOR METHOD



2. Under the Genomic Data Sharing (GDS) Policy, is NIH allowing investigators who are approved to download human datasets from NIH controlled-access repositories to use cloud computing?

In April 2015, NIH released the NIH Position Statement on Use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH Genomic Data Sharing Policy and is now allowing investigators to request permission to transfer controlled-access genomic and associated phenotypic data obtained from NIH designated repositories under the auspices of the GDS Policy to public or private closed systems for data storage and analysis. NIH expects cloud computing systems to meet the data use and security standards outlined in NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy as well as the institution's own IT security requirements and policies. Investigators who wish to use cloud computing for storage and analysis will need to indicate in their Data Access Request (DAR) that they are requesting permission to use cloud computing, identify the cloud service provider or providers that will be employed, and describe how the cloud computing service will be used to carry out their proposed research.

The NIH strongly recommends that investigators consult with institutional IT leaders, including the Chief Information Officer (CIO) and the institutional Information Systems Security Officer (ISSO) or equivalents to develop the formal information security plan prior to receipt of controlled access data from the NIH, and ***institutional signing officials should validate that an appropriate security plan is in place prior to accepting liability for data loss or breach on behalf of the institution***. This document provides an overview of security principles for data, access, and physical security to ensure confidentiality, privacy, and accessibility of data. This is a minimum set of requirements; additional restrictions may be needed by your institution and should be guided by the knowledge of the user community at your institution as well as your institution's IT requirements and policies.

Data on 150,000 patients exposed in another misconfigured AWS bucket

Patient Home Monitoring failed to lock down public access to its online server, exposing personal data of 150,000 patients.

By **Jessica Davis** | October 12, 2017 | 02:02 PM

Kromtech Security researchers have discovered yet another unsecured Amazon S3 bucket. This time, the cloud server in question was linked to HIPAA-covered entity, Patient Home Monitoring, a vendor that provides U.S. patients with disease management services and in-home monitoring.

A COMMUNITY EFFORT OF GENOMIC DATA PRIVACY PROTECTION

2014 – 2017 iDASH genomic data privacy and security protection competition http://www.humangenomeprivacy.org

IDASH PRIVACY WORKSHOPS*

* Supported by U54HL108460 initially, and then by R13HG009072

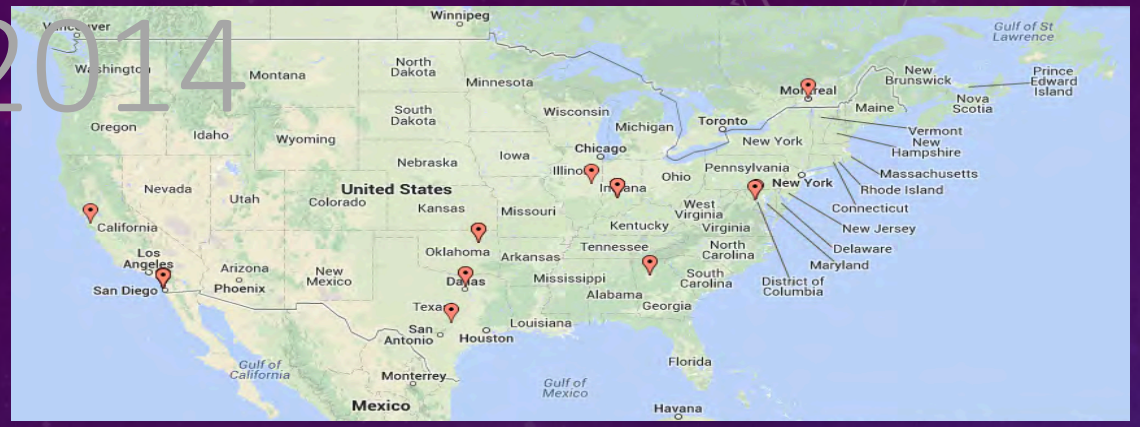
http://www.humangenomeprivacy.org/

An interdisciplinary challenge on genomic privacy research

- Motivated by real world biomedical applications and with participation of privacy technology experts, Biomedical researchers, ELSI researchers (academia and industry)
- Developed practical yet rigorous solutions for privacy preserving genomic data sharing and analysis
- Demonstrated feasibility of secure genome data analysis and dissemination using differential privacy, MPC, HE, SGX
- Reported in the media (e.g., Nature News)

http://www.nature.com/news/extreme-cryptography-paves-way-to-personalized-medicine-1.17174

2014



2 countries
9 states
33 registrations

- Privacy preserving data sharing
- Secure release of genome analysis results

2015

5 countries
7 states
50+ registrations

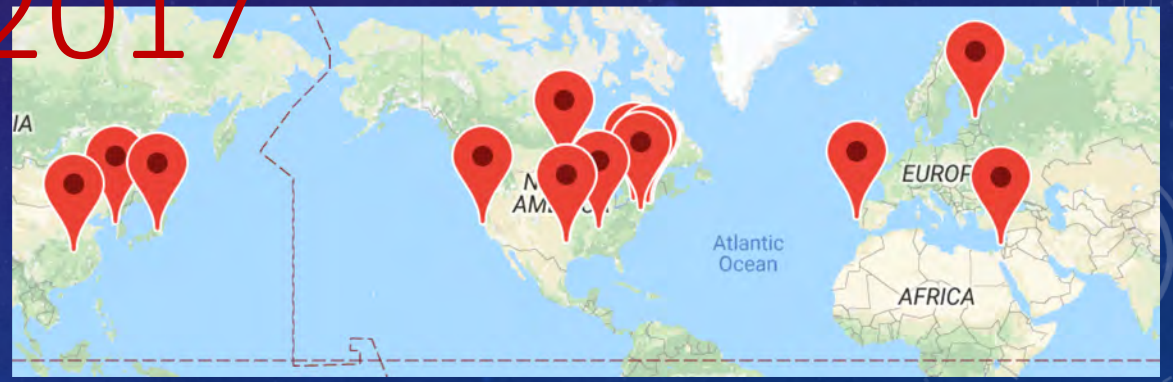
- Homomorphic Encryption for GWAS (MAF&Chi-Squared)
- Secure Collaboration on DNA Analysis

2016

13 countries
10+ states
75+ registrations

- Privacy-Preserving Search of Similar Cancer Patients across Organizations
- Testing for Genetic Diseases on homomorphically Encrypted Genomes
- Protecting queries in Beacon service

2017



19 countries
65+ Teams

- Secure Record De-duplication
- Secure GWAS using SGX
- Homomorphic logistic regression

APPLICATIONS ENABLED BY HOMOMORPHIC ENCRYPTION

Year	Homomorphic encryption Applications	Winning Team	Problem setup	Run time	Peak memory cost
2015	Minor Allele Frequency Chi-squared statistics	Stanford/MIT	610 SNPs and 200 individuals	1.847 (seconds)	13 (MB)
	Hamming Distance	IBM	100K sequences	472.2 (seconds)	2.168 (GB)
	Approximate Edit Distance	Microsoft Research	10K sequences	181.92 (Seconds)	1.295 (GB)
2016	Genetic testing	Microsoft Research	(1 query (1 variant) / 50 VCF files [100k])	59.58 (seconds)	83.6 (MB)
2017	Logistic Regression	Seoul National University	Datasets with 1422 records and 18 features	10.360 (minutes)	2775.333 (MB)

2018 IDASH COMPETITION

- New challenges will be announced soon in earlier April.
- Http://www.humangenomeprivacy.org