National Institute of General Medical Sciences

National Human Genome Research Institute

IDASH PRIVACY & SECURITY WORKSHOP 2020

*supported by NHGRI R13HG009072

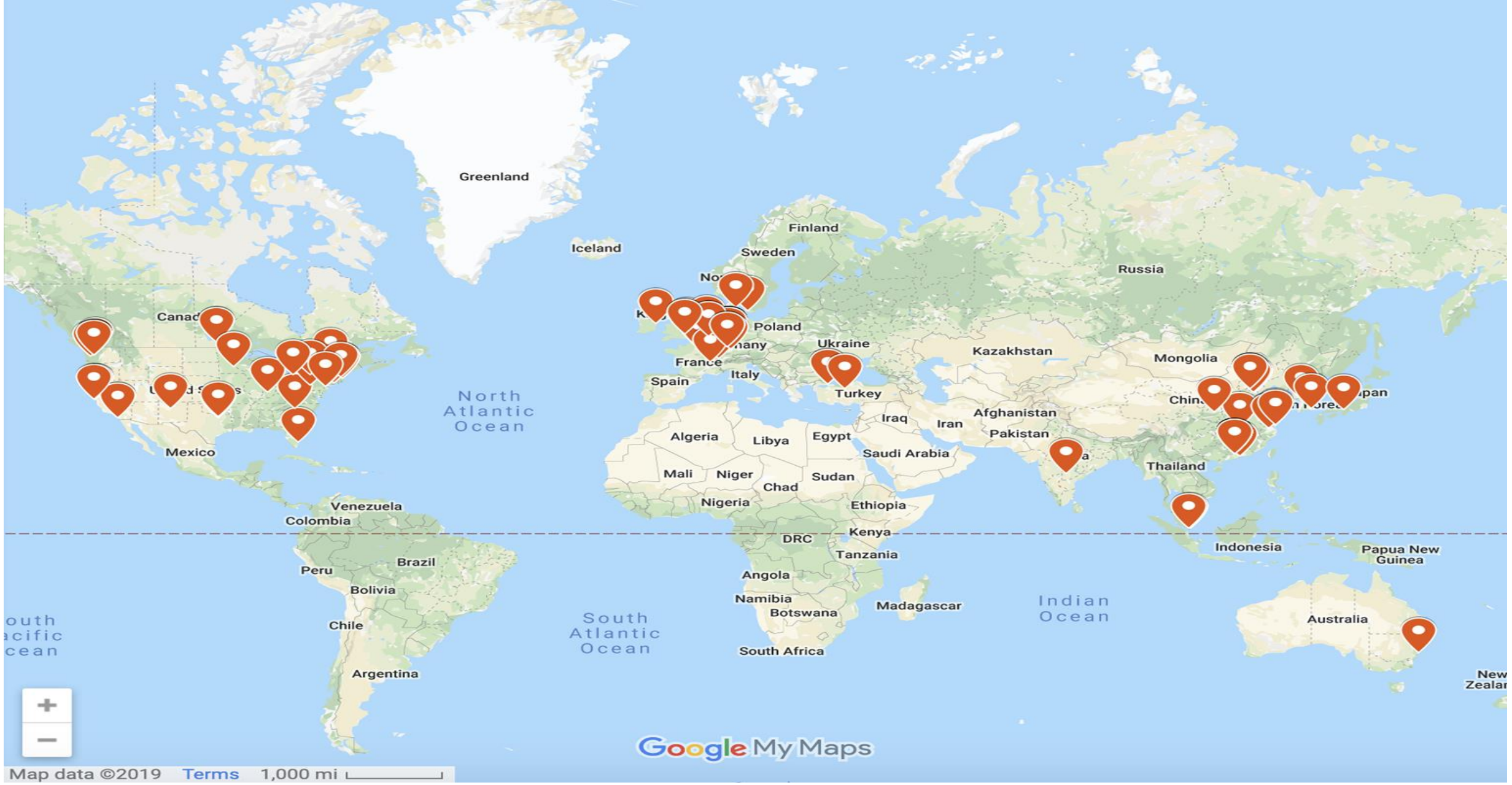
Organizers:

Lucila Ohno-Machado, Tsung-Ting Kuo (UCSD)

Track 1: Miran Kim (UNIST), Arif Harmanci (UTHEALTH), Xiaoqian Jiang (UTHEALTH),

Track 2 & 3: Haixu Tang, XiaoFeng Wang (IUB)

Setting the Stage



Four tracks
Three host institutions

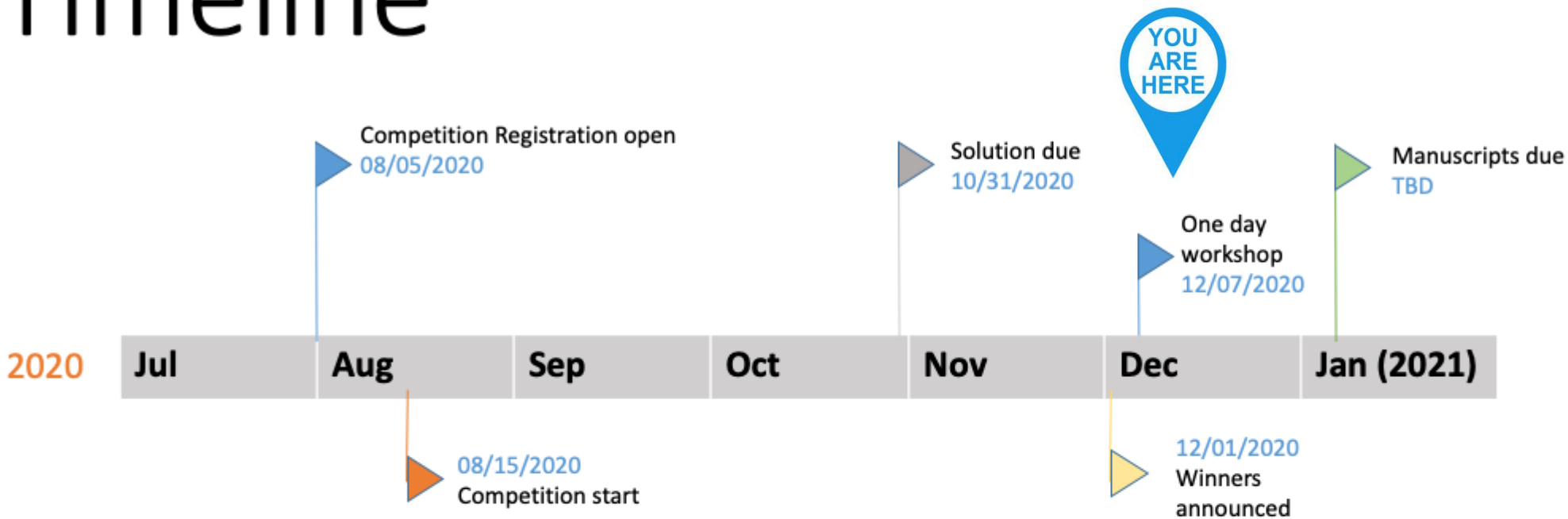
115 registered teams

21 countries
4 continents

5 months of preparation
Dozens of developers

1 month of evaluation
using ~100 VMs

Timeline



Registration 08/05/2020-08/15/2020

Develop solutions 08/15/2020-10/31/2020

Code evaluation by organizers 10/31/2020-12/01/2020

Manuscript preparation 12/01/2020-TBD

THREE TRACKS OF COMPETITION TASKS

- TRACK 1: Secure multi-label Tumor classification using Homomorphic Encryption
- TRACK 2: Privacy-preserving clustering of single-cell transcriptomics data in SGX
- TRACK 3: Differentially private federated learning for cancer prediction model

Acknowledgements

- **Track 1 (UTH)**
  - Luyao Chen
- **Track 2 and 3 (IUB)**
  - Weijie Liu, Tianhao Mao, Diyue Bu, Lei Wang
- **Workshop (UCSD)**
  - Morgan Von Ebke
- **Sponsors**
  - NIH/NHGRI

National Institute of General Medical Sciences

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

National Human Genome Research Institute

Track I: Secure multi-label Tumor classification using Homomorphic Encryption

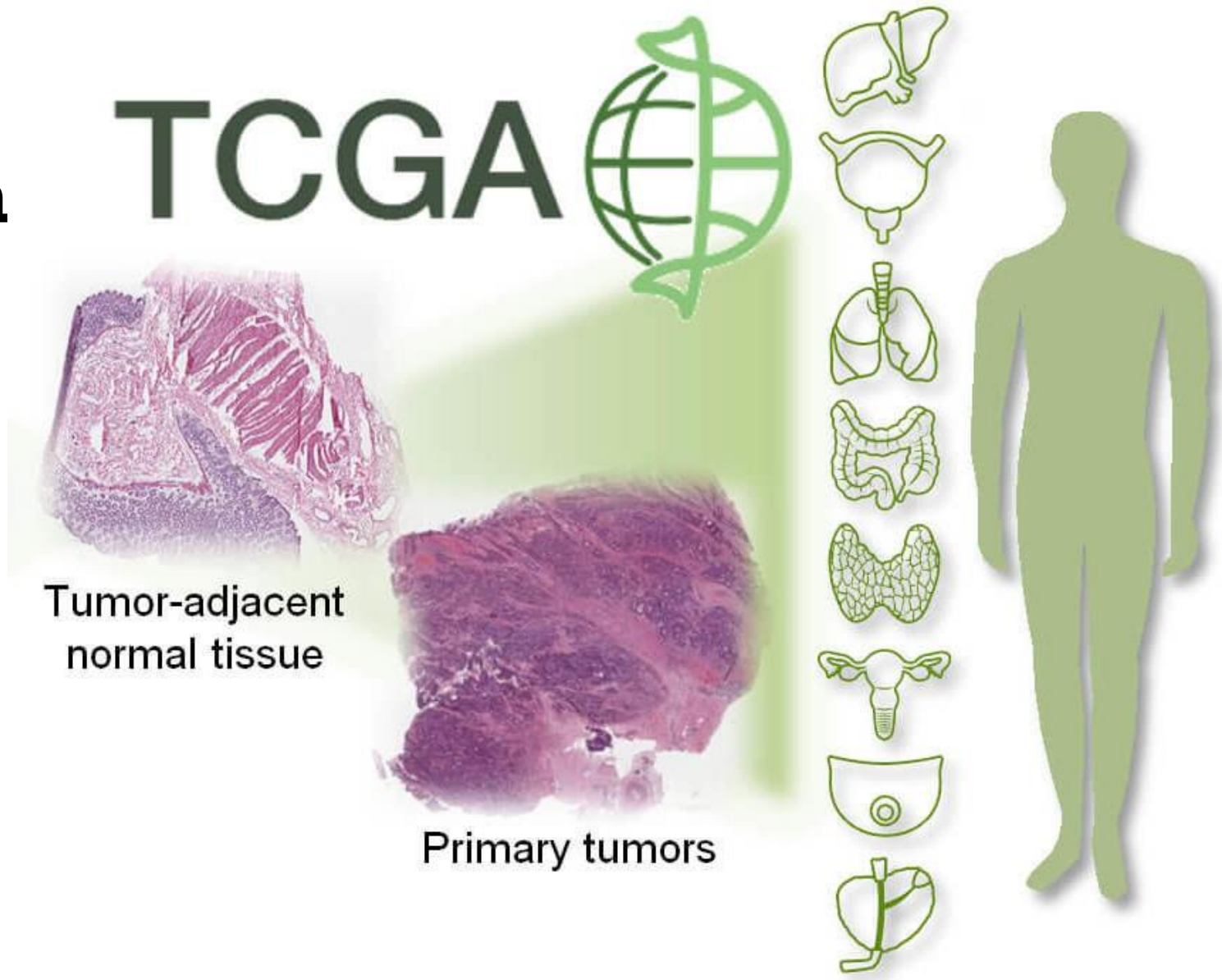
*supported by NHGRI R13HG009072

Xiaoqian Jiang, Luyao Chen, Miran Kim, Arif Harmanci

UT Health at Houston

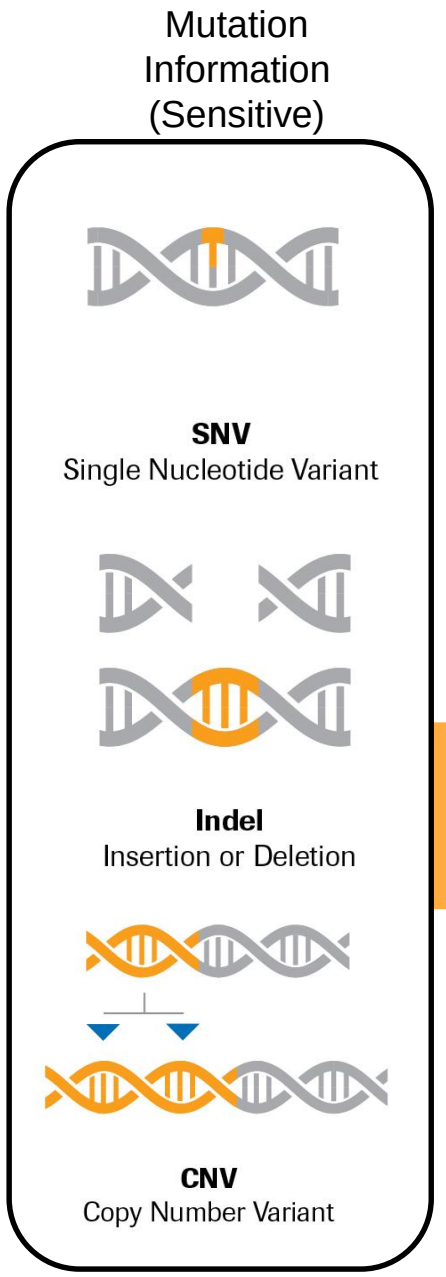
TRACK 2: Secure Outsourced Tumor Classification

- **Goal**: The goal of this track is to develop a secure tumor classification (type and position) using the mutation data for patients
- Why?
  - Increases understanding of genetic determinants of tumors biology
  - Identifies “Drivers” of tumors that originate in different tissues
  - Provides therapeutic targets for treatment of tumors from different origins
- Patient privacy becomes important while sharing the variant information for classifying the tumor with respect to molecular information



Track 1: Secure Outsourced tumor classification

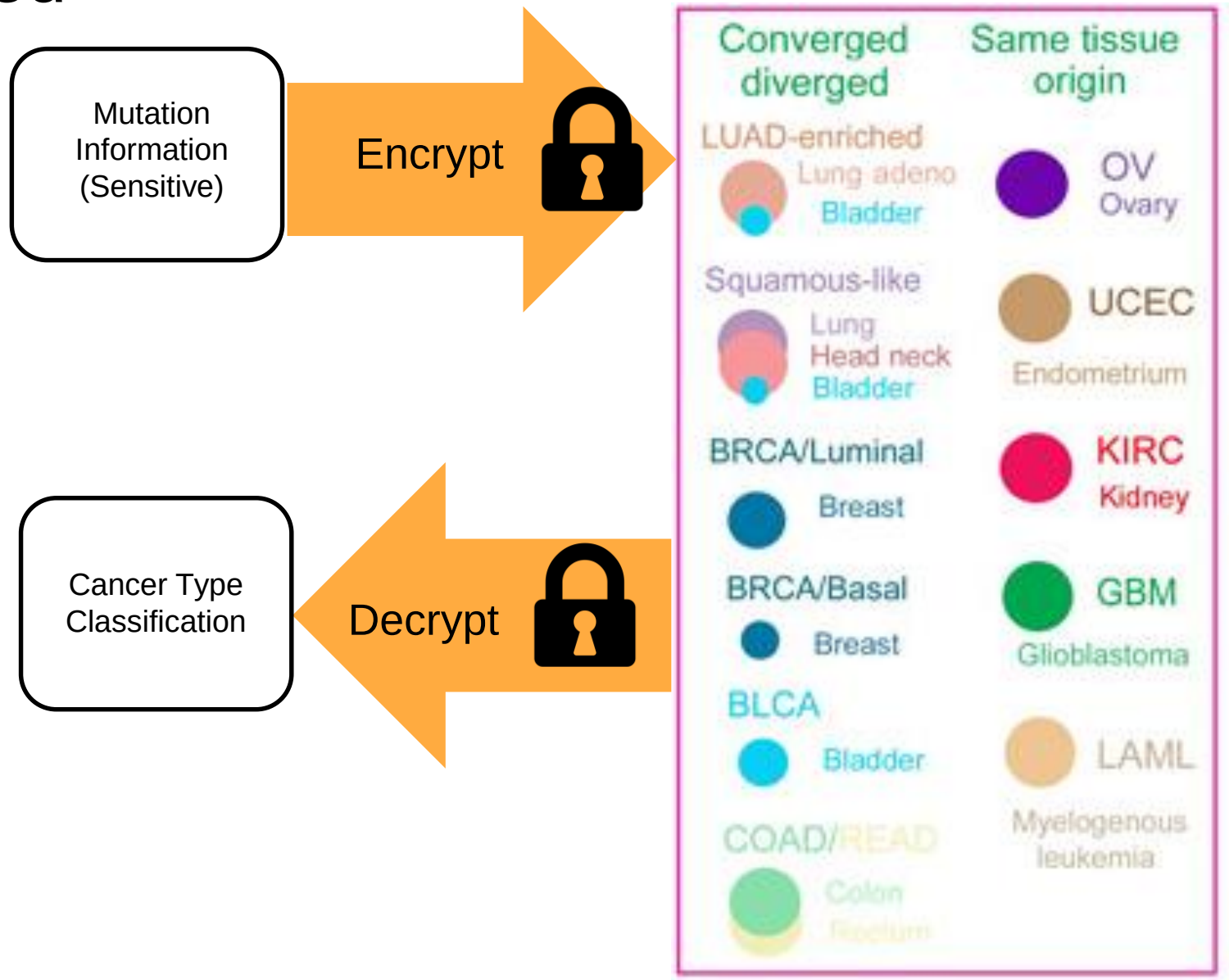
- **11 Cancer types from TCGA are used**
- Publicly available mutation information is used to generate the training and evaluation datasets
- Single nucleotide variants and copy number variants are used as inputs
  - Variants are heterogeneously distributed
  - Variant information is treated as the sensitive identifying information
- The main challenges are handling the heterogeneity of the variants
  - Develop an efficient feature matrix
- Manage the run time below 5 minutes run time for whole evaluation that contains 900 samples



Das, Abecasis, Browning, Annu. Rev. Genom. Hum. Genet., 2018.

Track 1: Secure Outsourced tumor classification

- **11 Cancer types from TCGA are used**
- Single nucleotide variants and copy number variants are used as inputs
  - Variants are heterogeneously distributed
  - Variant information is treated as the sensitive identifying information
- Each team is asked to build secure classification models using the mutation data
- The mutation data was preprocessed to build feature matrix then encrypted at the client side
- 2,713 samples in the training data
- 909 samples in the evaluation dataset



Das, Abecasis, Browning, Annu. Rev. Genom. Hum. Genet., 2018.

TRACK I: Teams and Evaluation

➡ ***36 participating teams***

➡ ***15 submitted their solutions***

➡ ***13 completed the evaluation process***

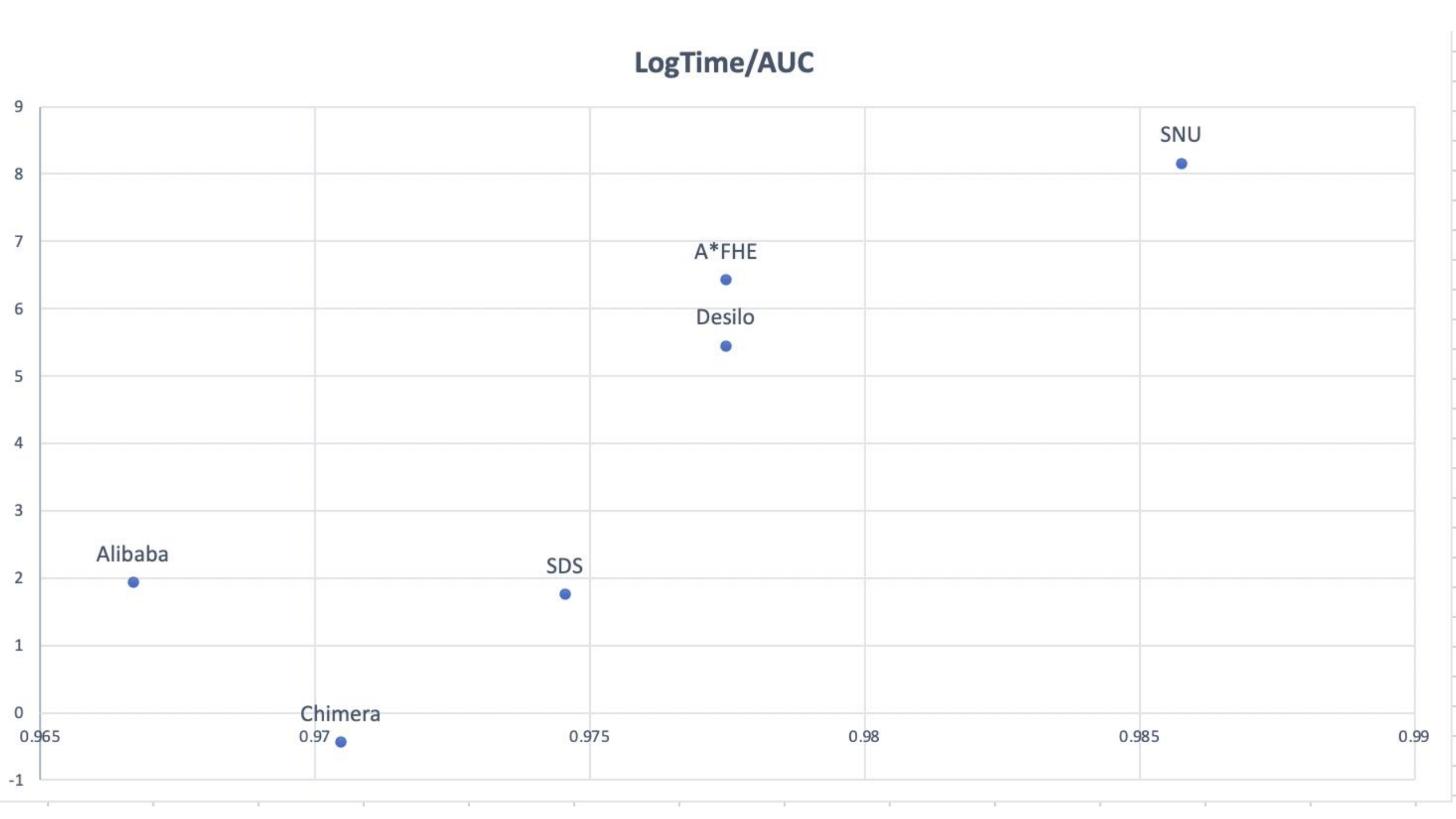
- Evaluation Environment and Platform
  - Centos 7 Server 2 * Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz (24 cores / 48 threads), 768G RAM
  - Each docker container was limited to access 32G RAM and 4 virtual CPUs
  - The actual CPU clock runs between 1.20G ~ 3.70G, so that we give some tolerance for the solutions slightly more than 5 minutes
  - The AUC is calculated based on the probability/score of prediction returned by each solution.
    - roc_auc_score function of sklearn is used for evaluation, average = 'micro' method is used

AUC Sorted Rankings

Team Name	Institution	Country	End-End Time (Sec.)	AUC
SNU	Seoul National University	South Korea	286.13	0.9857596265677173
A*FHE	A*STAR	Singapore	186.01	0.9774805423336614
Desilo	Desilo	South Korea	43.70	0.9774713444941853
SamsungSDS	SamsungSDS	South Korea	3.39	0.9745466735892511
Chimera	Inpher	Switzerland	0.74	0.9704708204593838
Alibaba Gemini Lab	Alibaba Group	China	3.82	0.9668795482408528
Rosetta	PlatON and Matrix Elements	China	13.58	0.9598181490316249
CodeHopper	Temple University	United States	239.79	0.9542025049589666
gencSU	Sabanci University	Turkey	16.01	0.9380109188036514
NYUAD-Yale	NYU Abu Dhabi / Yale University	UAE / USA	308.32	0.9286282148566892
Team Genigma	Sandia National Laboratories	USA	51.00	0.9218367601336591
Team Inspire	Georgia State University	United States	61.00	0.884644146482855
LangWolf	Wuhan University Of Technology	China	48mins	0.8607356940314493
ANT_SCI	Ant Group	China	N/A	N/A
BitSecure Group	Tsinghua University	China	N/A	N/A

End-End Time Sorted Rankings

Team Name	Institution	Country	End-End Time (Sec.)	AUC
Chimera	Inpher	Switzerland	0.74	0.9704708204593838
SamsungSDS	SamsungSDS	South Korea	3.39	0.9745466735892511
Alibaba Gemini Lab	Alibaba Group	China	3.82	0.9668795482408528
Rosetta	PlatON and Matrix Elements	China	13.58	0.9598181490316249
gencSU	Sabanci University	Turkey	16.01	0.9380109188036514
Desilo	Desilo	South Korea	43.70	0.9774713444941853
Team Genigma	Sandia National Laboratories	USA	51.00	0.9218367601336591
Team Inspire	Georgia State University	United States	61.00	0.884644146482855
A*FHE	A*STAR	Singapore	186.01	0.9774805423336614
CodeHopper	Temple University	United States	239.79	0.9542025049589666
SNU	Seoul National University	South Korea	286.13	0.9857596265677173
NYUAD-Yale	NYU Abu Dhabi / Yale University	UAE / USA	308.32	0.9286282148566892
LangWolf	Wuhan University Of Technology	China	48mins	0.8607356940314493
ANT_SCI	Ant Group	China	N/A	N/A
BitSecure Group	Tsinghua University	China	N/A	N/A


LogTime/AUC
9
8
7
6
5
4
3
2
1
0
-1
0.965
0.97
0.975
0.98
0.985
0.99
SNU
A*FHE
Desilo
Alibaba
SDS
Chimera

Final Rankings (Considering Time and AUC)

1st Place: SNU, Desilo, Chimera, SamsungSDS

2nd : Alibaba Gemini Lab, A*FHE

Team Name	Institution	Country	End-End Time (Sec.)	AUC
SNU	Seoul National University	South Korea	286.13	0.9857596265677173
Desilo	Desilo	South Korea	43.70	0.9774713444941853
Chimera	Inpher	Switzerland	0.74	0.9704708204593838
SamsungSDS	SamsungSDS	South Korea	3.39	0.9745466735892511
Alibaba Gemini Lab	Alibaba Group	China	3.82	0.9668795482408528
A*FHE	A*STAR	Singapore	186.01	0.9774805423336614

National Institute of General Medical Sciences

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

National Human Genome Research Institute

Track II: Privacy-preserving clustering of single-cell transcriptomics data in SGX

*supported by NHGRI R13HG009072

Haixu Tang and XiaoFeng Wang

Indiana University at Bloomington

TRACK II: BACKGROUND

- **Background**: Single-cell RNA-seq technologies have advanced rapidly. Un-supervised learning methods such as dimension reduction and clustering algorithms are now widely used to group cells of the same type or subtypes based on the gene expression profiles of hundreds to thousands of single cells e.g., from tumor or normal tissues.
- **Confidential Computing**: Trusted Executive Environment (TEE, e.g., Intel's SGX) provides an ideal infrastructure for hosting such privacy-preserving analyses of single-cell transcriptomics data by a data user.
    - Only the computing task (i.e., clustering of single-cell gene expression profile) approved by the owner of the input data is allowed be performed on the data;
    - The data user does not see the content of input data, which are encrypted in a way that only the client and the TEE can decrypt.
- The **purpose** of this task is to test the efficiency of an unsupervised clustering algorithm in SGX when applied to massive single-cell RNA-seq data.

TRACK II: CHALLENGE

- We challenge participating teams to implement a given clustering algorithm (CIDR) on the Intel SGX platform, so the algorithm can operate inside the SGX enclave.

- The implementation should protect both the input data: that is, any input, intermediate and output data should be encrypted outside the enclave.

- We do not consider side channel leaks in this task.

Lin, P., Troup, M. & Ho, J.W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol 18, 59 (2017).

TRACK II: EVALUATION

- **Experimental setting:** We provide a testing dataset along with the plain implementation of the expected clustering algorithm. *Each team is challenged to implement the model under SGX*. For this purpose, the team is allowed to develop approximation algorithms so it can work in the enclave, as long as its accuracy is largely preserved and privacy is fully protected (except side channel leaks). The testing dataset will be used to evaluate the implemented model. The solution may utilize the computational resource outside the enclave, including the CPU, memory and hard disk, as long as all the data and the model are fully protected (encrypted at least 128-bit security level). The submitted solution cannot involve any addition party. Pre-computing time will be measured as part of the performance overhead.

- **Evaluation Criteria:** All submissions should meet security requirements (**at least 128-bit security level**). Also we expect that the protected model mostly preserves the accuracy of the original model but will also compare different models' performance when their accuracy comes close.

National Institute of General Medical Sciences

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING

National Human Genome Research Institute

Track III: Differentially private federated learning for cancer prediction model

*supported by NHGRI R13HG009072

Haixu Tang and XiaoFeng Wang

Indiana University at Bloomington

TRACK III: BACKGROUND

- **Background:** Training a modern machine learning model often requires a large amount of data distributed across multiple organizations. Oftentimes, however, data owners could be reluctant to share their data (e.g., genome data from human subjects), even in the encrypted form, due to the restrictions of their organizational privacy policies. Therefore, it becomes highly desired to allow two or more owners to build a joint ML model while ensuring the privacy protection of the data and the policy compliance.

- The **purpose** of this task is to understand the feasibility of building such a machine learning model among multiple collaborative parties so that the data of each organization does not leave its premise and the information (e.g., intermediate parameters of the model) exchanged across the parties during the computation is properly protected under differential privacy.

TRACK III: CHALLENGE

- We challenge participating teams to implement a federate learning algorithm that can be trained jointly by two parties each holding their individual training datasets, i.e., gene expression levels on a group of patients with known phenotypes (disease or not). The implementation should not share the input data but can exchange intermediate results (e.g., intermediate model parameters) with the other party during the training process; however, noise should be added to the shared intermediate results to ensure differential privacy under a given budget, across the whole learning process. Each team can choose any ML algorithm to accomplish the task.

TRACK III: EVALUATION

- **Experimental setting:** We anticipated solutions containing two programs for the training and testing purpose, respectively. The training program may take as input a dataset of gene expression profiles with the same format as the given testing dataset, and output a model file (containing model parameters, etc, in the format recognizable by the testing program). The testing program takes as input the model file and makes prediction for a given input instance (i.e., a column of gene expression profile). The training program should operate on two separate machines that communicate intermediate results (model parameters) during training. The testing program is a standalone program running on a single computer.

- **Evaluation Criteria:** Submissions are qualified if they meet the differential privacy requirement under a given privacy budget. Qualified solutions will be ranked based on their performances, including their prediction accuracy., total running time, and the communication cost (the rounds and sizes of data exchange). The evaluation team will run the training code on the released data for up to 24 hours. The solutions that do not complete within 24 hours will be disqualified.

TRACK II: Teams Completing the Task

Among 23 participating teams, 3 completed the task (listed alphabetically)

Team (Affiliation)
Angel PowerFL (Tencent)
Morse (Ant Financial Services Group) *not functional*
ZMCTeam (Zhejiang University)

TRACK II: BEST-PERFORMING TEAMS

Evaluated on random 3,000/5,000/10,000 cells of dataset GSE131907.

Team	Affiliation	Dataset (# of cells)	Running time	Adjusted Rand Index (ARI)	Baseline ARI
Angel PowerFL	Tencent	3,000	5m26.739s	0.6097527	0.3883327
		5,000	22m29.150s	0.5909339	0.6482618
		10,000	4h22m32.912s	0.5752788	0.5859439
ZMCTeam	Zhejiang University, City University of Hong Kong	3,000	25m44.776s	0.3883327	0.3883327
		5,000	1h35m10.977s	0.6482618	0.6482618
		10,000	18h3m44.112s	0.5859439	0.5859439

Kim, N., Kim, H.K., Lee, K., Hong, Y., Cho, J.H., Choi, J.W., Lee, J.I., Suh, Y.L., Ku, B.M., Eum, H.H. and Choi, S., 2020. Nature communications.

TRACK III: Teams Completing the Task

Among 55 participating teams, 16 completed the task (listed alphabetically)

Team (Affiliation)	Team (Affiliation)
A*FHE (A*STAR)	FLR (Owkin)
Angel PowerFL (Tencent)	Manticore (Inpher)
BiuBiuBiu (Ant Group)	Morse (Ant Group)
CodeHopper (Temple University)	PrivCom (Baidu)
Consensys Health (ConsenSys Health)	RucPriv (Renmin University of China)
DP-FL (Purdue University)	Samsung SDS Research (Samsung SDS)
DSP (University of Manitoba)	Team GersteinLab (Yale University)
FederBoost (Zhejiang University)	ZJU-DPFL (Zhejiang University)

TRACK III: EVALUATION PERFORMANCE

Evaluated on random 117 records of dataset BC-TCGA.

Team	# of machines	Accuracy	Running time (training+test)	DP Criteria
Inpher Manticore	2	100%	0.682219s+0.71s	ε=3, δ=0
DP-FL	2	99.1%	1.180s+0.625s	ε=3, δ=0
FLR (Owkin)	2	99.1%	2.258s+0.603s	ε=3, δ=1e-05
DSP	2	99.1%	10.840s+2.670s	ε=3, δ=0
Angel PowerFL	2	100%	36.1239s+1.122s	ε=3, δ=1e-05
PrivCom	2	100%	1m24.430s+2.428s	ε=3, δ=1e-05
Morse	2	94.87%	2m48.588s+1.993s	ε=3, δ=0
Samsung SDS Research	2	89.1%	13.474s+11.747s	ε=2, δ=1e-05
Team GersteinLab	2	85.47%	4m3.28s+5.165s	ε=1, δ=0
BiuBiuBiu	2	89.74%	1m14.314s+6.387s	ε=10, δ=1e-05

Haozhe Xie, Jie Li, Qiaosheng Zhang, and Yadong Wang, 2016. Computational biology and chemistry.

TRACK III: EVALUATION PERFORMANCE

Evaluated on random 117 records of dataset BC-TCGA.

Team	# of machines	Accuracy	Running time (training+test)	DP Criteria
A*FHE	1	98.29%	19.44s+7.42s	ε=3, δ=0.1
FederBoost	1	100%	5m54.991+2.370s	ε=5, δ=0
CodeHopper	1	100%	6m58.854s	ε=5.38, δ=1e-03
RucPriv	1	98.2%	17m44.948s+4.985s	ε=4, δ=0
ZJU-DPFL	1	96.6%	15.995s+1.9s	ε=5, δ=1e-05
ConsenSys Health	1	95%	19m8.359s+3m13.156s	ε=1.05, δ=1e-05

Haozhe Xie, Jie Li, Qiaosheng Zhang, and Yadong Wang, 2016. Computational biology and chemistry.

Winning Teams

	Team	Member(s)
1st Place	**Angel PowerFL**	Yao Zhang (Tencent Blade Team) Chen Hou (Tencent Angel PowerFL Team) Huanran Xue (Tencent Angel PowerFL Team) Zhikai Chen (Tencent Blade Team) Huaming Rao (Tencent Angel PowerFL Team) Bo Zhang (Tencent Blade Team) Yong Cheng (Tencent Angel PowerFL Team) Yangyu Tao (Tencent Angel PowerFL Team)
2nd Place	**ZMCTeam**	Yuan Chen, Zhejiang University Leqian Zheng, City University of Hong Kong Cong Wang, City University of Hong Kong Yajin Zhou, Zhejiang University Jianfeng Zhu, Hangzhou MiShu Technology Kui Ren, Zhejiang University

Winning Teams

	Team	Member(s)
1st Place	**Inpher Manticore**	Sergiu Carpov, Nicolas Gama- Mariya Georgieva (Inpher, Lausanne, Switzerland).
2nd Place	**DP-FL**	Ninghui Li, Purdue University Zitao Li, Purdue University (work done during internship at Alibaba Group) Tianhao Wang, Purdue University Bolin Ding,, Alibaba Group
3rd Place	**FLR (Owkin)**	Constance Beguier, Jean Ogier du Terrail, Iqraa Meah, Mathieu Andreux, Eric W. Tramel (Owkin)
3rd Place	**DSP**	Md Momin Al Aziz, Monowar Anjum and Noman Mohammed Data Security and Privacy lab, Computer Science, University of Manitoba