National Cybersecurity Center of Excellence

NCCoE Virtual Workshop on Cybersecurity of Genomic Data

Wednesday, January 26, 2022, 11:00 AM – 4:30 PM (ET)

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Challenges from the Field: Research Perspective

Jean-Pierre Hubaux
(EPFL/ Global Alliance for Genomics and Health [GA4GH])

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Protecting and Sharing Genomic Data: a Swiss/European Perspective

Prof. Jean-Pierre Hubaux, EPFL
Co-Founder of Tune Insight SA

Work done in close collaboration with Lausanne University Hospital (CHUV)

With gratitude to all the colleagues I have had the privilege to work with

NIST National Institute of Standards and Technology U.S. Department of Commerce

NCCoE NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

About Switzerland

8.5 inhabitants

26 cantons (states), each with its own laws

Most of the health system managed by the cantons, not the federal government; the latter defines the overall policy and strategy

Data protection laws very similar to EU GDPR

Very strong political decentralization

One of highest GDP/capita in the world

Strong pharma: Roche, Novartis,…

5 university hospitals

2 federal institutes of technology: EPFL (Lausanne), ETH (Zurich)

Use case for Swiss Personalized Oncology Project: federated analytics platform for research and molecular tumor board

Q1: How many adult cancer patients consenting on reuse of routine data for research with diagnosis of a malignancy on or after 1st January 2015, mutations in BRAF gene and under anti-PD-1 are there?

Explore

Q2: Among these patients, what is the overall survival for patients with and without a mutation on position 600 of the BRAF gene?

Analysis

Universitätsspital Basel

USZ Universitäts Spital Zürich

INSELSPITAL

CHUV

HUG

SPHN Swiss Personalized Health Network

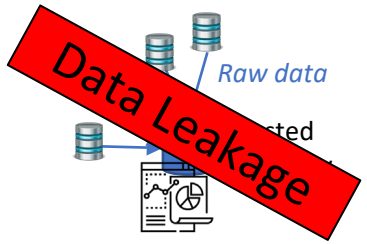
Swiss Personalized Oncology

The Main Challenges we Faced

- Multi-disciplinary nature of the problem: bio-informaticians, clinicians, geneticists, hospital IT specialists, hospital lawyers, data protection authorities, ethicists, computer scientists
- Mess of the health data
- Financial sustainability of the solution

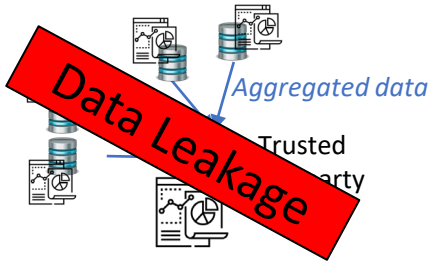
Distributed Learning - Current Approaches

a) Fully centralized



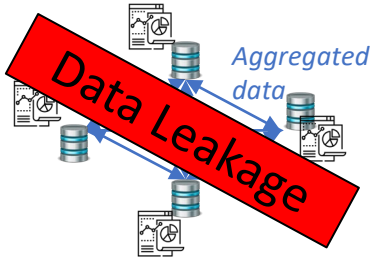
Examples:
- All of Us
- EGA
- Genomics England

Meta-analysis



https://covidclinical.net/

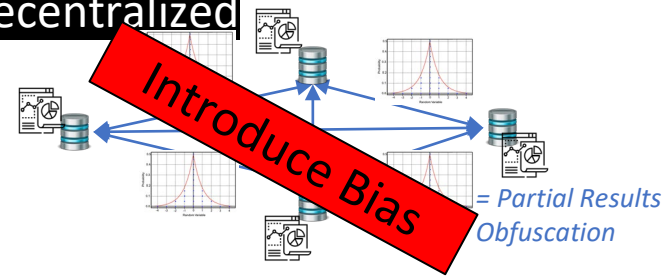
Decentralized



http://www.datashield.ac.uk

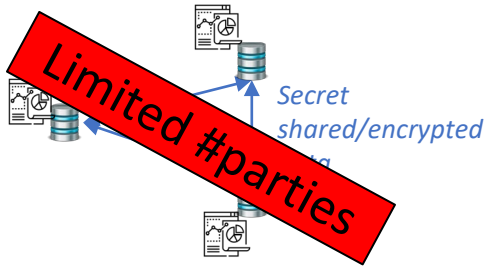
Personalized Health Train (PHT)

Differential Privacy Decentralized



Examples:
- M. Kim et al. "Secure and Differentially Private Logistic Regression for Horizontally Distributed Data," TIFS 2019
- M. Abadi et al. Deep learning with differential privacy. In ACM CCS, 2016.
- Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In NIPS, 2009.

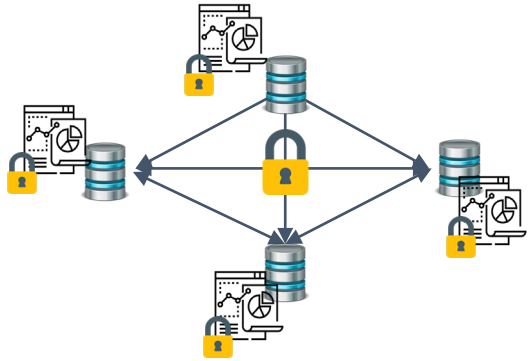
(e) Cryptographic (SMC, HE) Decentralized



Examples:
- A. Gascón et al.. Privacy-preserving distributed linear regression on high-dimensional data. PETS, 2017.
- P. Mohassel and Y. Zhang. SecureML: A system for scalable privacy-preserving machine learning. In IEEE S&P, 2017.

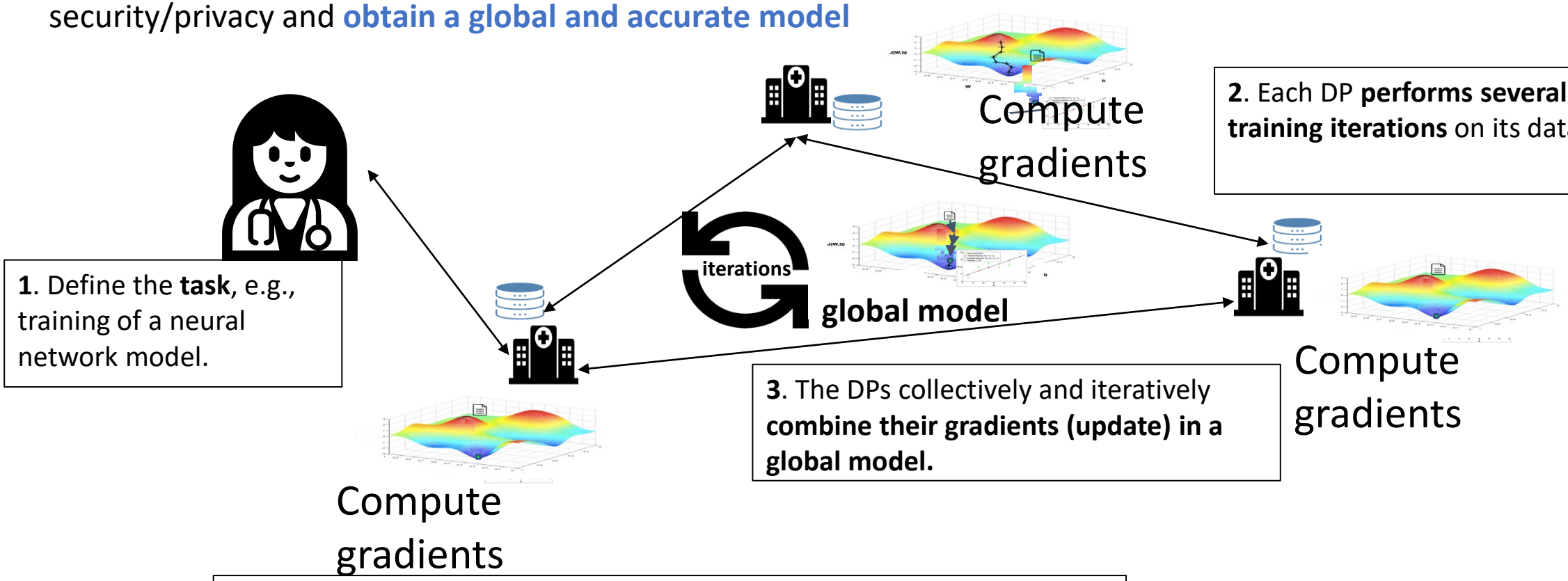
Our approach

→ **Data + Model Confidentiality as long as 1 entity is honest**
→ No data outsourcing
→ Scales with #parties
→ Exact results



Privacy-Preserving Federated Neural Network Learning

Solution: The data providers (DPs) collaborate to enable a joint gradient descent while protecting their security/privacy and **obtain a global and accurate model**



S.Sav, A. Pyrgelis, J.R. Troncoso-Pastoriza, D. Froelicher, J.P. Bossuat , J.S. Sousa and J.P. Hubaux,

POSEIDON: Privacy-Preserving Federated Neural Network Learning. NDSS, 2021

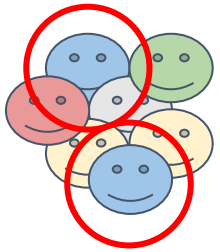
MedCo

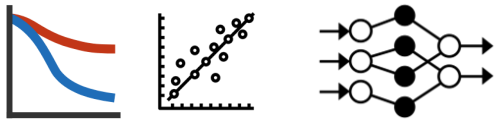
- **Distributed software platform** for federated cohort exploration and analytics of clinical and genomic data
- Co-developed by EPFL and CHUV
- Built on top of the i2b2 cohort explorer (i2b2 is used by 250+ hospitals worldwide)

- Relies on **advanced cryptographic techniques**
  → Multi-party homomorphic encryption (MHE)

- Code-reviewed and pen-tested by third-party industrial companies, compliant with hospitals' information security policies

- Main functionalities
  - **MedCo-Explore: cohort exploration**
    - Obtaining cohort sizes for clinical research studies based on inclusion/exclusion criteria
  - **MedCo-Analysis: federated analytics**
    - Survival analysis
    - ML training and testing





April 2020: MedCo deployed at 3 hospitals

MedCO
Collective protection of medical data

SPHN Swiss Personalized Health Network

Strategic Focus Area
Personalized Health and Related Technologies

HUG CHUV INSEL

EPFL software to enable secure data-sharing for hospitals

The MedCo system aims to facilitate medical research on pathologies – such as cancer and infectious diseases – by enabling secure computations on decentralized data. The unique software has recently been deployed at three Swiss hospitals.

02.04.20

LINKS

- MedCo
- LDS
- Video

- First application:
  Swiss Personalized Oncology project:
  → melanoma data and beyond
- Planned deployment at Zurich University Hospital
- Ongoing international deployments: USA, NL, Italy, France

Data Protection Impact Assessment (DPIA) for multisite medical data analysis (June 2021)

Centralized approach with standard pseudonymization

Threat	Threat likelihood	Threat impact	Risk	Risk level
Unlawful access to the system	Unlikely	High	Loss of data confidentiality	Moderate
Malicious use of the system	Possible	High	Loss of data confidentiality	High
Loss of data	Unlikely	Minor	Loss of data integrity, data unavailability	Minor
Data leak of host/cloud	Possible	High	Loss of data confidentiality	High
Collusion of host/cloud	Possible	High	Loss of data confidentiality	High
Corrupted or malicious host/cloud	Possible	High	Data unavailability, loss of data integrity, loss of data confidentiality, loss of data correctness	High
Unavailability of host/cloud	Possible	Minor	Data unavailability, loss of data correctness	Moderate
Re-identification/attribute inference	Possible	High	Loss of data confidentiality	High

Federated approach enhanced **with MedCo**

Threat	Measure introduced with MedCo	Threat likelihood	Threat Impact	Risk	Risk level
Unlawful access to the system	1	Unlikely	Minor	Loss of data confidentiality	Low
Malicious use of the system	1, 2, 4, 10	Possible	Minor	Loss of data confidentiality	Low
Loss of data	3, 5	Unlikely	Minor	Loss of data integrity, data unavailability	Low
Data leak	4, 5, 8, 9, 10	Unlikely	Minor	Loss of data confidentiality	Low
Collusion between nodes	4, 9	Unlikely	Moderate	Loss of data confidentiality	Moderate
Corrupted or malicious nodes	2, 5, 6, 7, 8, 9	Unlikely	Moderate	Data unavailability, loss of data integrity, loss of data confidentiality, loss of data correctness	Moderate
Unavailability of of nodes	6, 7	Possible	Minor	Data unavailability, loss of data correctness	Moderate
Re-identification or attribute inference	1, 2, 4, 9, 10	Unlikely	Minor	Loss of data confidentiality	Low

Feedback from Swiss authorities on MedCo DPIA

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Data Protection and Information Commissioner

"... the threat impact of most risks with the MedCo system shows to be clearly lower than with traditional systems. Since data processed within the Medco framework remain encrypted during computation, an attacker would cause little damage. **As no entity has the full decryption key, it seems indeed unlikely that he could decrypt and abuse the stolen data**. ..."

13 September 2021

GDPR legal compliance: partial aggregates are not personal data anymore, they are anonymous

JMIR Publications
Advancing Digital Health & Open Science

Articles | Search articles

Journal of Medical Internet Research | Journal Information | Browse Journal | Subm

Published on 25.2.2021 in Vol 23, No 2 (2021): February

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/25120, first published October 19, 2020.

Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis

James Scheibner [1,2]; Jean Louis Raisaro [3,4]; Juan Ramón Troncoso-Pastoriza [5]; Marcello Ienca [1]; Jacques Fellay [3,6,7]; Effy Vayena [1]; Jean-Pierre Hubaux [5]

Article | Open Access | Published: 11 October 2021

Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption

David Froelicher, Juan R. Troncoso-Pastoriza, Jean Louis Raisaro, Michel A. Cuendet, Joao Sa Sousa, Hyunghoon Cho, Bonnie Berger, Jacques Fellay & Jean-Pierre Hubaux

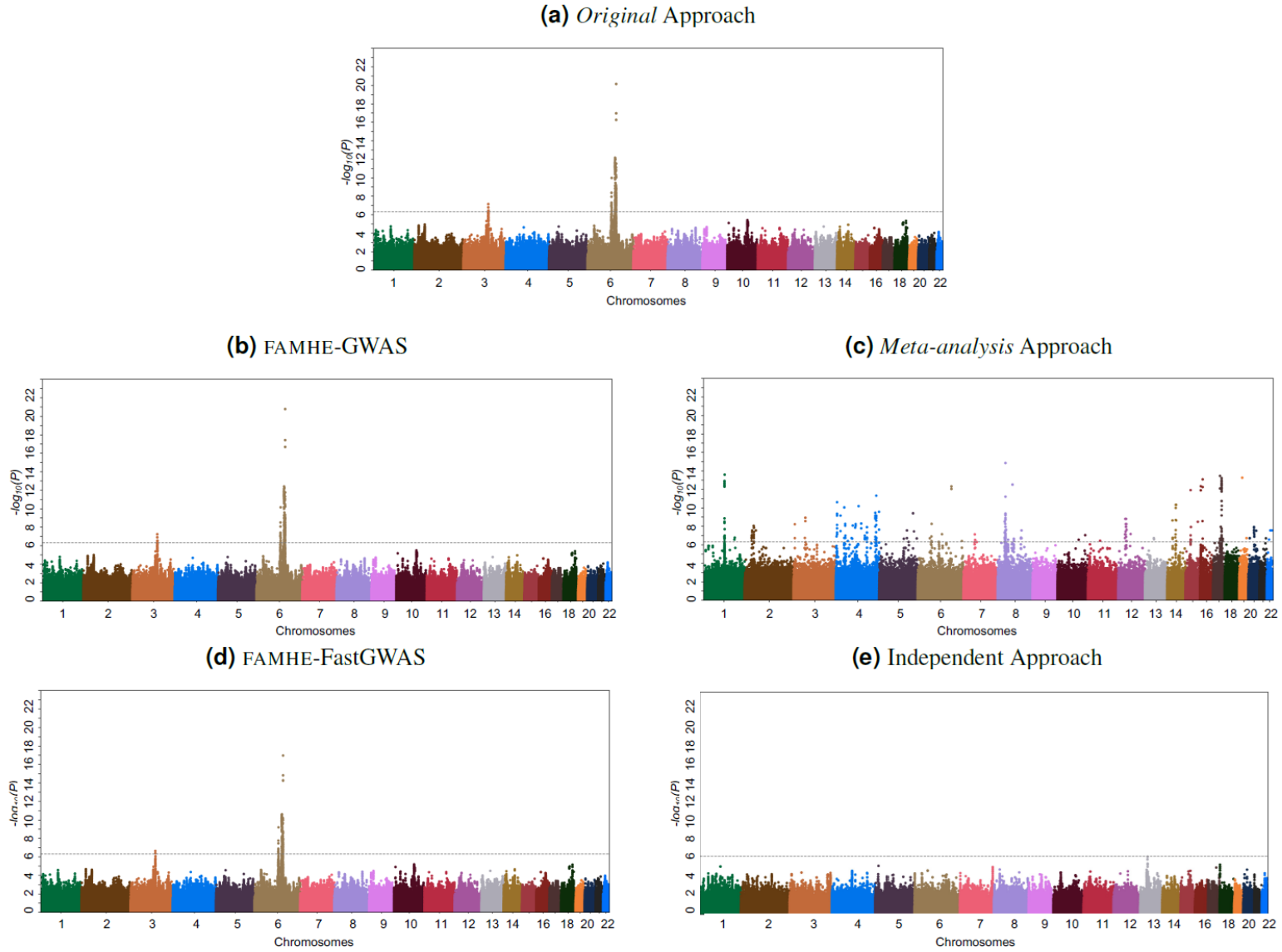
Nature Communications **12**, Article number: 5910 (2021) | Cite this article

Metrics

Abstract

Using real-world evidence in biomedical research, an indispensable complement to clinical trials, requires access to large quantities of patient data that are typically held separately by multiple healthcare institutions. We propose FAMHE, a novel federated analytics system that, based on multiparty homomorphic encryption (MHE), enables privacy-preserving analyses of distributed datasets by yielding highly accurate results without revealing any intermediate data. We demonstrate the applicability of FAMHE to essential biomedical analysis tasks, including Kaplan-Meier survival analysis in oncology and genome-wide

FAMHE: Privacy-Preserving Federated Analytics for Precision Medicine with MHE - GWAS



[Original approach] McLaren, P. J. et al. Polymorphisms of Large Effect Explain the Majority of the Host Genetic Contribution to Variation of HIV-1 Virus Load. Proc. Natl. Acad. Sci. 112, 14658–14663 (2015).

[FAMHE] Froelicher et al. Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption.

FAMHE: Genome-wide association study

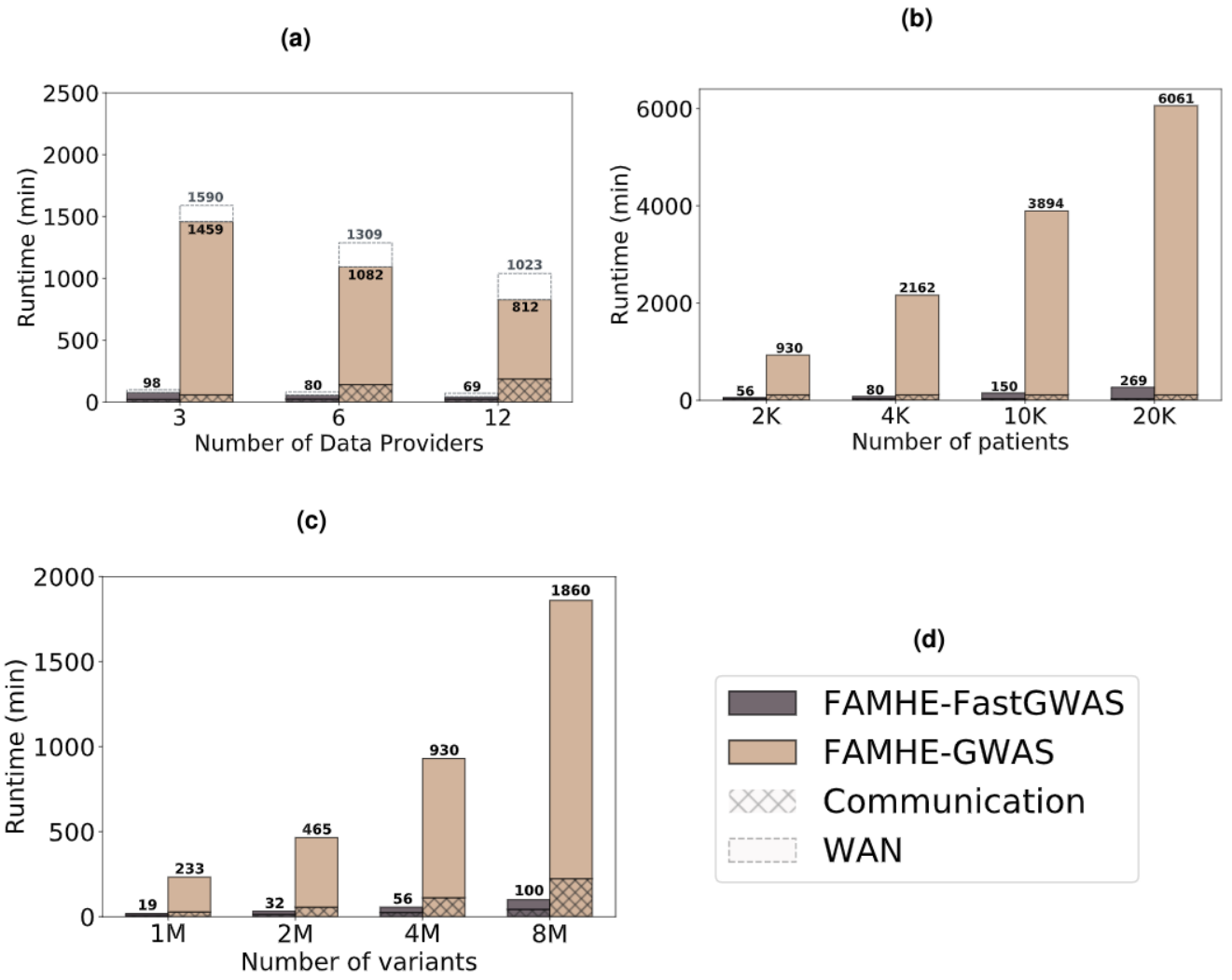
Default: 1857 patients spread among 12 data providers.

→ scale in all dimensions

a. *With the number of data providers*

b. *With the number of patients*

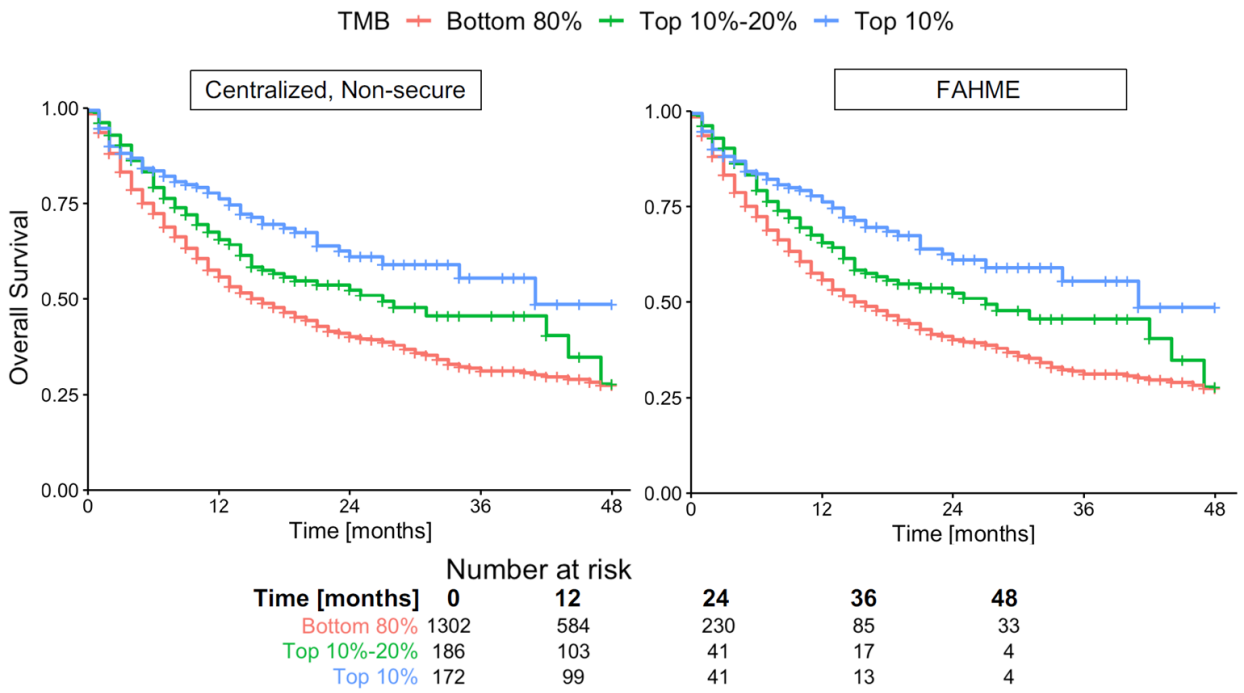
c. *With the number of variants*

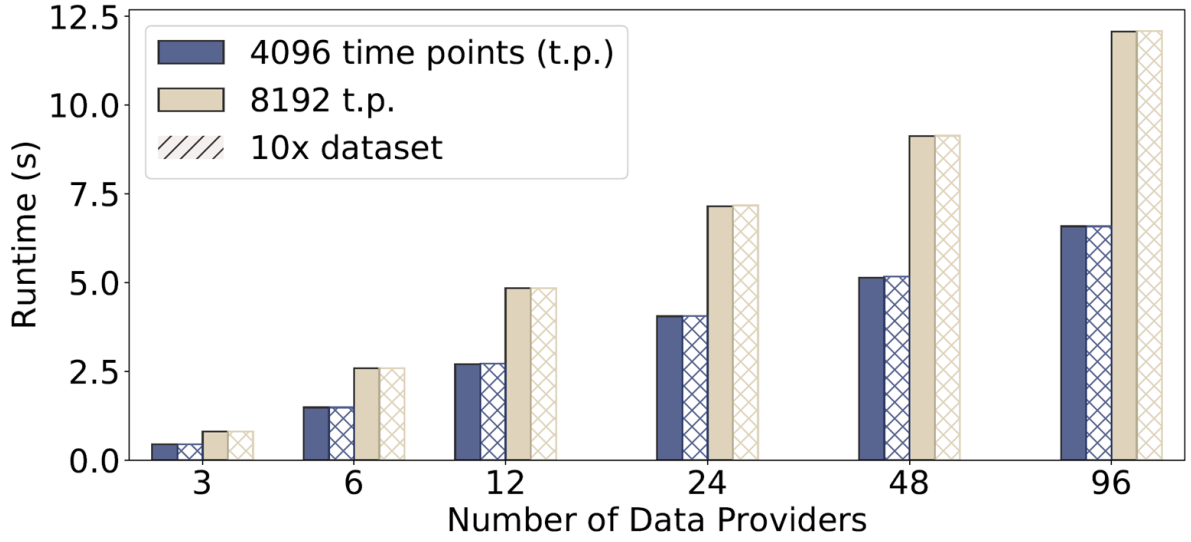


[Centralized] McLaren, P. J. et al. Polymorphisms of Large Effect Explain the Majority of the Host Genetic Contribution to Variation of HIV-1 Virus Load. Proc. Natl. Acad. Sci. 112, 14658–14663 (2015).
[FAMHE] Froelicher et al. Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption. Submitted to Nat. comms. https://doi.org/10.1101/2021.02.24.432489

FAMHE: Privacy-Preserving Federated Analytics for Precision Medicine with MHE - Survival curves (Kaplan-Meier)

Data split among 3 data providers:

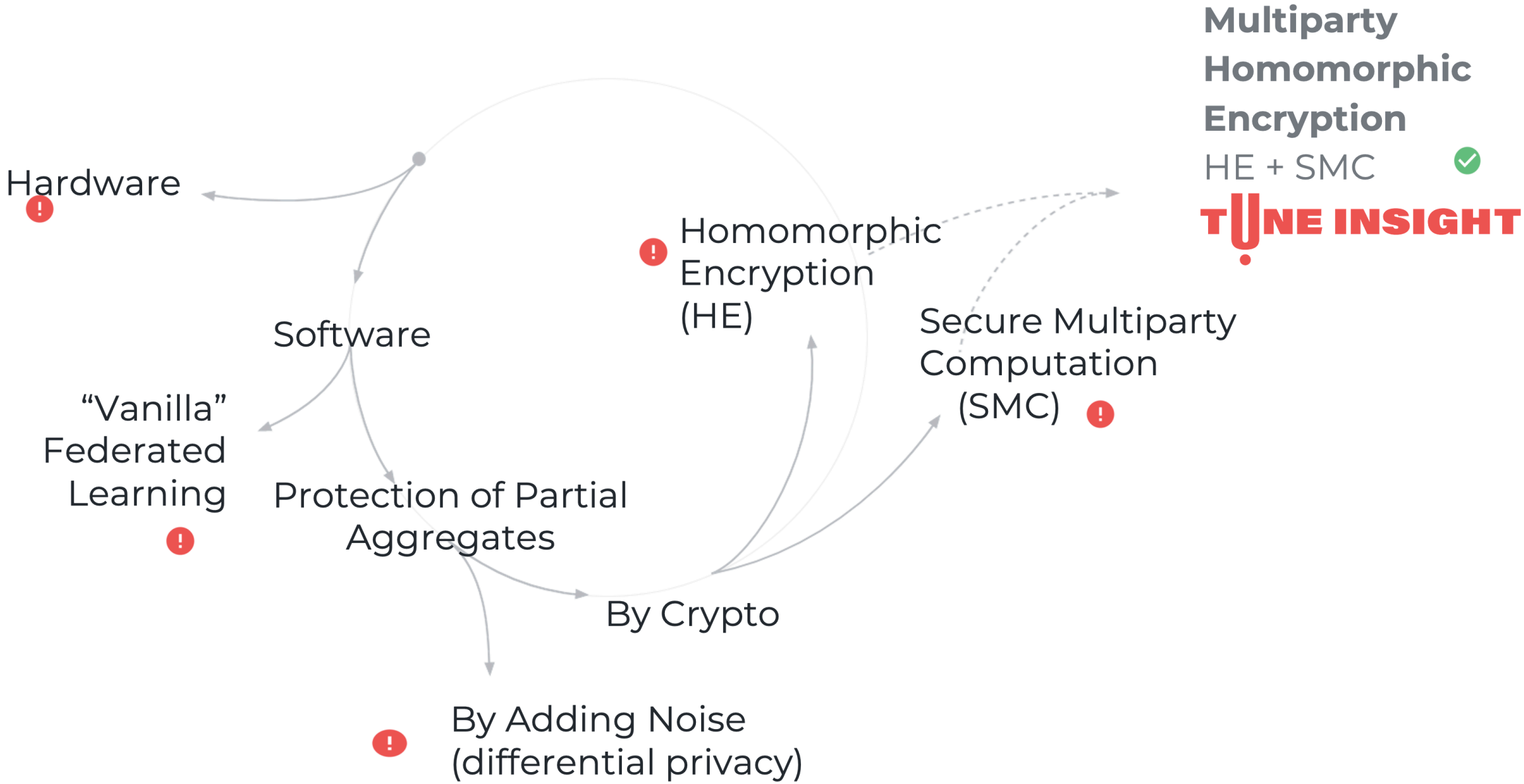




[Centralized] Samstein, R. M. et al. Tumor Mutational Load Predicts Survival after Immunotherapy across Multiple Cancer Types. Nat. genetics 51, 202–206 (2019).

[FAMHE] Froelicher et al. Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption.

Share without Sharing: Available Options



Enterprise Data & Analytics

$100B Market

(GlobalData, Enterprise Data and Analytics Market, 2020))

56% say own data is not enough, expanding ability to source external data

(Forrester, The Insights Professional's Guide To External Data Sourcing, 2019)

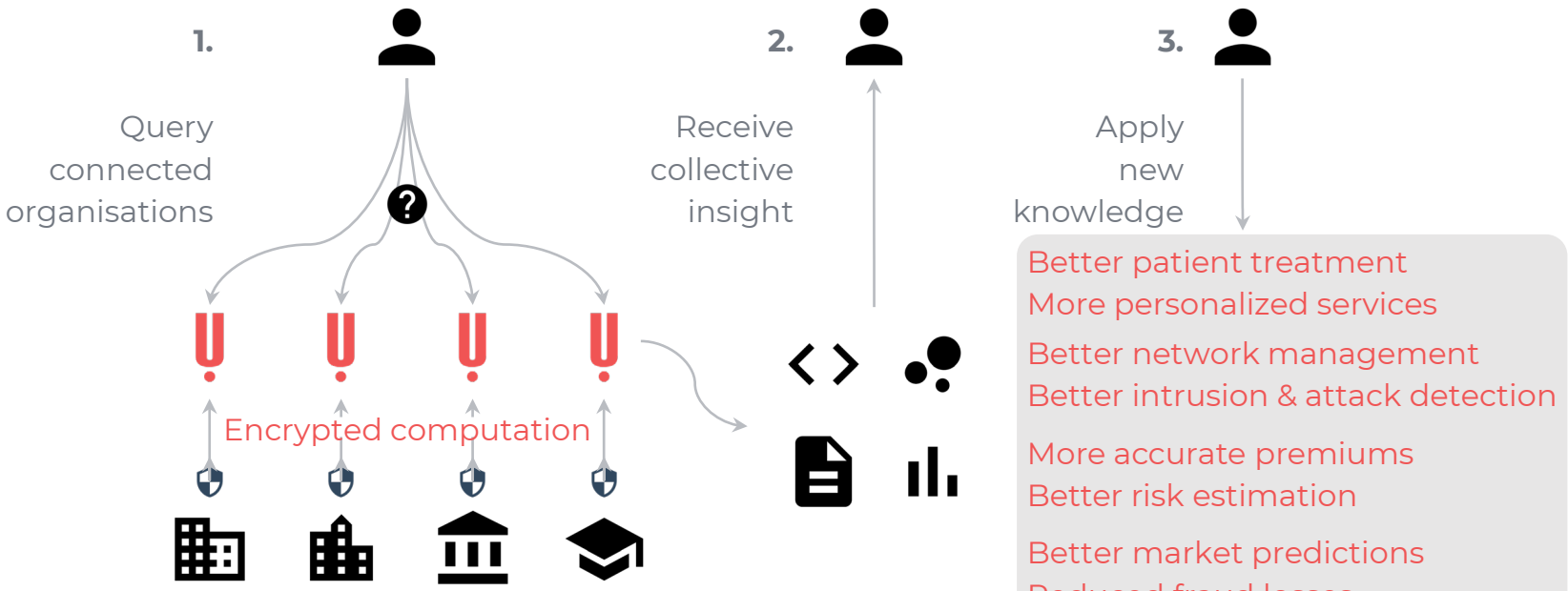
However, organizations are **prevented** to enter valuable data collaborations due to fear of **data leaks** and **data protection regulations**

TUNE INSIGHT

juan@tuneinsight.com

Cross-vertical enterprise SaaS enabling organizations to make better decisions, together, by orchestrating secure collaborations around their sensitive data.

- CHF400k in customer-paid projects including with Swiss Re, Armasuisse
- Pilot deployed at Swiss hospitals
- CHF100k EPFL Innogrant
- State-of-the-art post-quantum encryption technology
- Raised pre-seed with Wingman Ventures



Access to insights **Immediacy** **Compliance**

Personalization **Scalability** **Control**

MHE: mathematical proofs instead of vendor lock-in and side-channel attacks

	Software-based solutions (MHE)	Hardware-based solutions (e.g., Intel SGX)
System and trust model	**Decentralized** (federated computing, edge computing) or **centralized** (outsourced) systems	**Only centralized** systems (data has to be transferred to the TEE)
Assumptions	Protection against passive adversaries with quantum computing power: **processing infrastructure (including side-channels) and other data providers**	Protection against passive adversaries (other tenants); **limited protection against the processing infrastructure**; protection against side-channels is implementation-dependent
Implementation cost	**Tailored solution**; application-specific design; composition of cryptographic building blocks; limited range of efficient functionalities	**Available SDKs**; relatively easy conversion to secure enclave; general-purpose solutions; limited libraries and memory inside the enclave
Performance and overhead	**Less than 10x** overhead when full packing capacity is utilized (federated training of GLMs and NNs). Up to 4-5 orders of magnitude overhead for non-optimized or non-packed solutions	**Negligible** overhead for **regular instructions**; **4x overhead for memory** copy operations; **35x overhead for syscalls** to/from enclave
Response to newly discovered vulnerabilities	**Software patch** with protocol update; usually, no re-encryption of the data is needed	**Firmware** patch with variable **performance impact** (1x to 20x slow-down); **architecture change and hardware replacement**; **enclave code update** (update signatures, keys, and require new attestation)

GLM : Generalized Linear Model
MHE : Multi-party homomorphic encryption
NN : Neural Network
SDK : Software Development Kit
SGX : Software Guard eXtensions
TEE : Trusted Execution Environment

International collaborations

- Prof. Xiaoqian Jiang, UT Health
- GA4GH Data Security Work Stream
- MedCo now part of the i2b2 official community projects
- Prof. Shawn Murphy, HMS, and the ACT Network
- Broad Inst. + MIT
- Cancer Institute of the Netherlands
- …

Events devoted to the topic

- GenoPri.org: International workshop on genome privacy and security
  - Yearly workshop, typically co-located with GA4GH main annual event
- iDash - http://www.humangenomeprivacy.org/2021
  - Annual event with technical challenges on genome data protection and sharing

Conclusion

- We have solved the problem of GDPR-compliant federated learning for medical data, including genomic data
- Solution: Multi-party homomorphic encryption (MHE)
  - Perform computations without “seeing” the data
  - Rely on decentralized trust and mathematical proofs
  - No need to transfer the data
- Scalability with the number of data providers and the size of the datasets
- Green light from the Swiss federal data protection authority
- Support and development of new features: provided by Tune Insight

Contact me at jean-pierre.hubaux@epfl.ch

More information at https://medco.epfl.ch

Challenges from the Field: Individual's Perspective

John Verdi (Future of Privacy Forum)

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Privacy Risks and Challenges from the Perspective of Individual Rights

John Verdi, Senior Vice President of Policy at the Future of Privacy Forum.

National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Privacy Risks and Challenges from the Perspective of Individual Rights

FPF Work:

- In July 2018, the Future of Privacy Forum released Privacy Best Practices for Consumer Genetic Testing Services
- FPF developed the Best Practices following consultation with technical experts, regulators, leading consumer genetic and personal genomic testing companies, and civil society
- On January 1, 2022, California's Genetic Information Privacy Act (GIPA) became effective, codifying many of FPF's best practices

Privacy Risks and Challenges from the Perspective of Individual Rights

FPF Requirements:

- Express consent for collection, use, and retention of genetic data;
- Separate express consent for transfer of to third parties and for incompatible uses;
- Informed consent for research;
- Educational resources about the basics, risks, benefits, and limitations of genetic and personal genomic testing;
- Access, correction, and deletion rights;
- Valid legal process for disclosure to the government and transparency reports;
- Ban on sharing genetic data with third parties (such as employers, insurance companies, educational institutions, and government agencies) without consent or as required by law;
- Restrictions on marketing based on genetic data; and
- Strong data security protections and privacy by design

Privacy Risks and Challenges from the Perspective of Individual Rights

Privacy Risks and Challenges of genomic data:

- Unique, immutable biometric
- Potentially reveals information about identity
- Potentially reveals information about heritage
- Potentially reveals information about health
- Potentially reveals information about relatives' identities, heritage, and health
- Difficult or impossible to de-identify without undermining utility

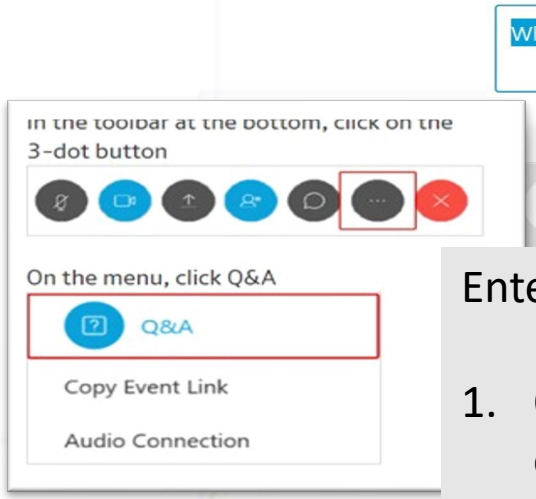
Privacy Risks and Challenges from the Perspective of Individual Rights

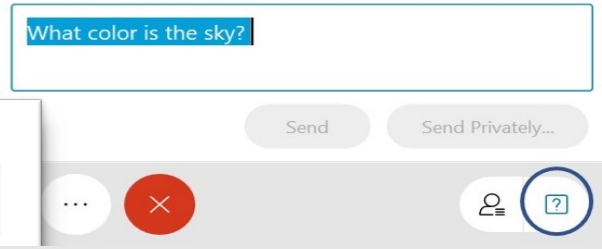
Privacy Risks and Challenges of genomic data:

- False identifications in criminal matters (evidence mishandling)
- Unexpected family connections and non-connections
- Dept. of Defense warning re: health tests and readiness reporting
- False identifications in criminal matters (remote relatives)
- Data breaches – e.g. 2020 GED Match law enforcement breach
- Re-identification attacks, e.g. cross-referencing clinical, research, and publicly available data sets

Challenges from the Field: Research and Individual's Perspectives

Moderated Questions and Answers





Enter your question in the Q&A panel.

1. On the right side, click on Q&A header to open the Q&A panel.
2. Type in the box **your name, organization and question**.
3. Click send.