Third ITU/WHO workshop on "Artificial Intelligence for Health"
Session 3: Benchmarking and Security
January 21, 2019, EPFL

DPPH
Data Protection in Personalized Health

Secure and Privacy-Preserving Data Sharing for Artificial Intelligence in Health

Jean-Pierre Hubaux
EPFL

With gratitude to the biomedical and CS researchers I have the privilege to work with

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

CHUV

Growing Concern: Medical Data Breaches

Around 5 declared breaches per week, each affecting 500+ people

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

Under Investigation | Archive | Help for Consumers

As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals. The following breaches have been reported to the Secretary:

Cases Currently Under Investigation

This page lists all breaches reported within the last 24 months that are currently under investigation by the Office for Civil Rights.

Show Advanced Options

Breach Report Results

Expand All	Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information
	Ohio Living	OH	Healthcare Provider	6510	09/07/2018	Hacking/IT Incident	Email
	Rockdale Blackhawk, LLC d/b/a Little River Healthcare	TX	Healthcare Provider	1494	09/07/2018	Unauthorized Access/Disclosure	Electronic Medical Record, Other
	J.A. Stokes Ltd.	NV	Healthcare Provider	3200	09/05/2018	Hacking/IT Incident	Desktop Computer, Electronic Medical Record, Network Server
	Reliable Respiratory	MA	Healthcare Provider	21311	09/01/2018	Hacking/IT Incident	Email
	Port City Operating Company doing business as St. Joseph's Medical Center	CA	Healthcare Provider	4984	08/31/2018	Loss	Other Portable Electronic Device
	Carpenters Benefit Funds of Philadelphia	PA	Health Plan	20015	08/31/2018	Hacking/IT Incident	Email

Technologies for Privacy and Security Protection

Traditional Encryption

- Protects data at rest and in transit
- Cannot protect computation

Homomorphic Encryption

- Protects computation in untrusted environments
- Limited versatility vs efficiency

Secure Multiparty Computation

- Protects computation in distributed environments
- High communication overhead

Trusted Execution Environments

- Protects computation with Hardware Trusted Element
- Requires trust in the manufacturer, vulnerable to side-channels

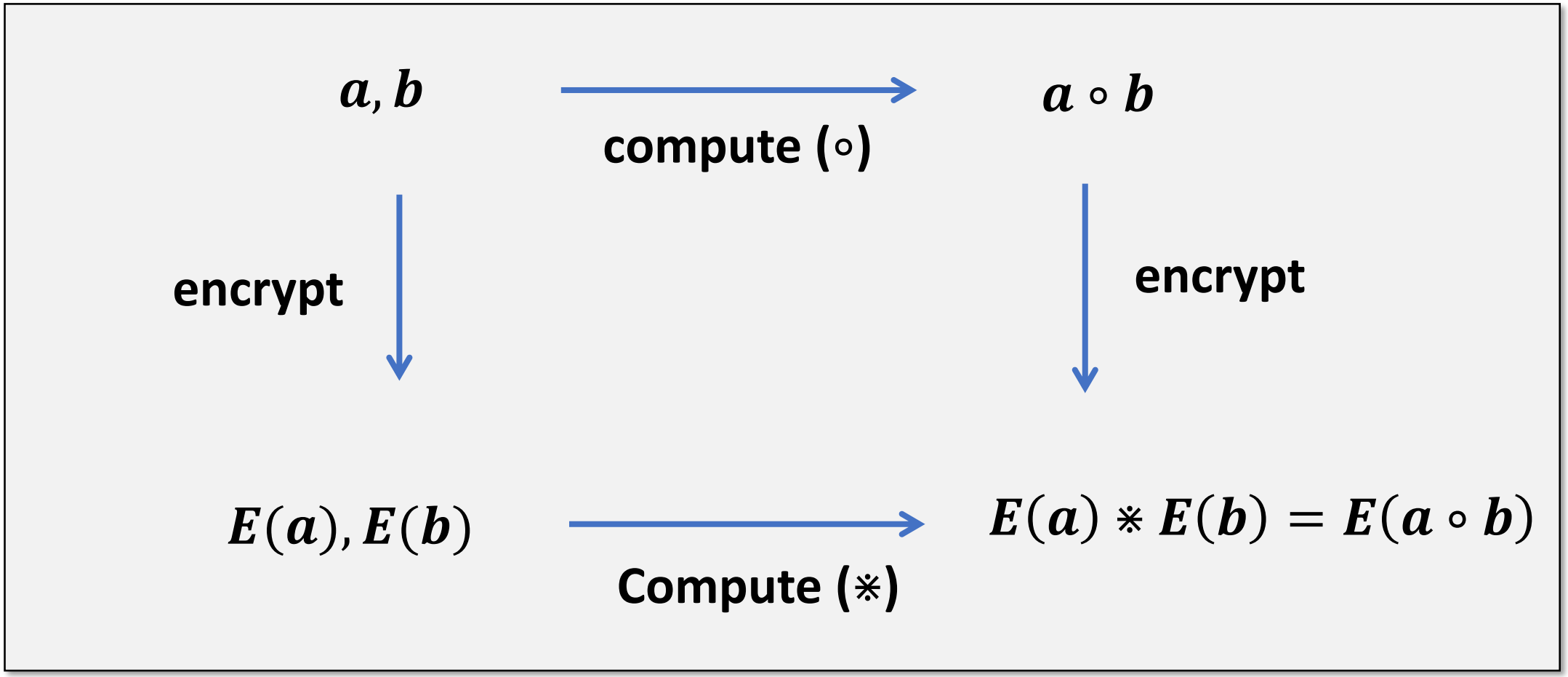
Differential Privacy

- Protects released data from inferences
- Degrades data utility (privacy-utility tradeoff)

Distributed Ledger Technologies (Blockchains)

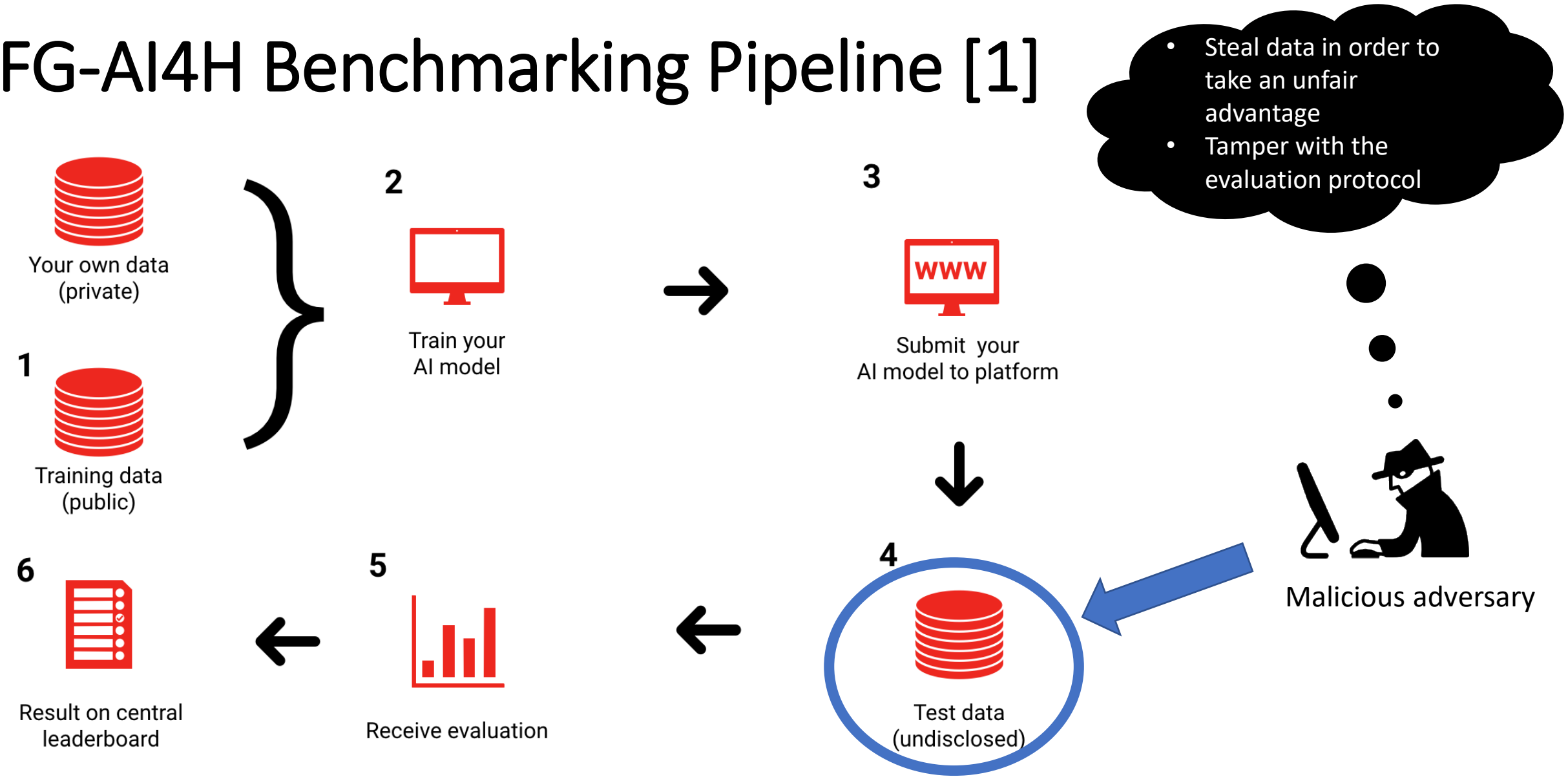
- Strong accountability and traceability in distributed environments
- Usually no data privacy

Homomorphic Encryption



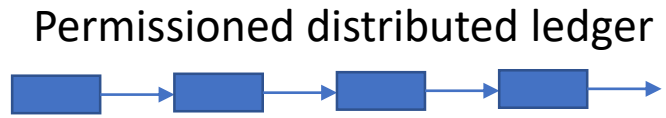
Homomorphic encryption enables computations directly on encrypted data.

FG-AI4H Benchmarking Pipeline [1]



[1] Salathé M, Wiegand T, Wenzel M and Kishnamurthy R, *Focus Group on Artificial Intelligence for Health*, White paper https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf

Privacy-Preserving Approach



- **Trust is distributed** within the collective authority => no single point of trust/failure
- Test data **confidentiality** is protected end-to-end with **collective homomorphic encryption**
- **Accountability** and **transparency** are provided by the use of **permissioned distributed ledger** where all actions are immutably logged

DPPH – Data Protection in Personalized Health

- 5 research groups across the ETH domain + SDSC (Swiss Data Science Center)
- Funding: 3 Millions CHFrs
- Duration: 3 years (4/2018 - 3/2021)
- Funding Program: ETH PHRT (Personalized Health and Related Technologies)

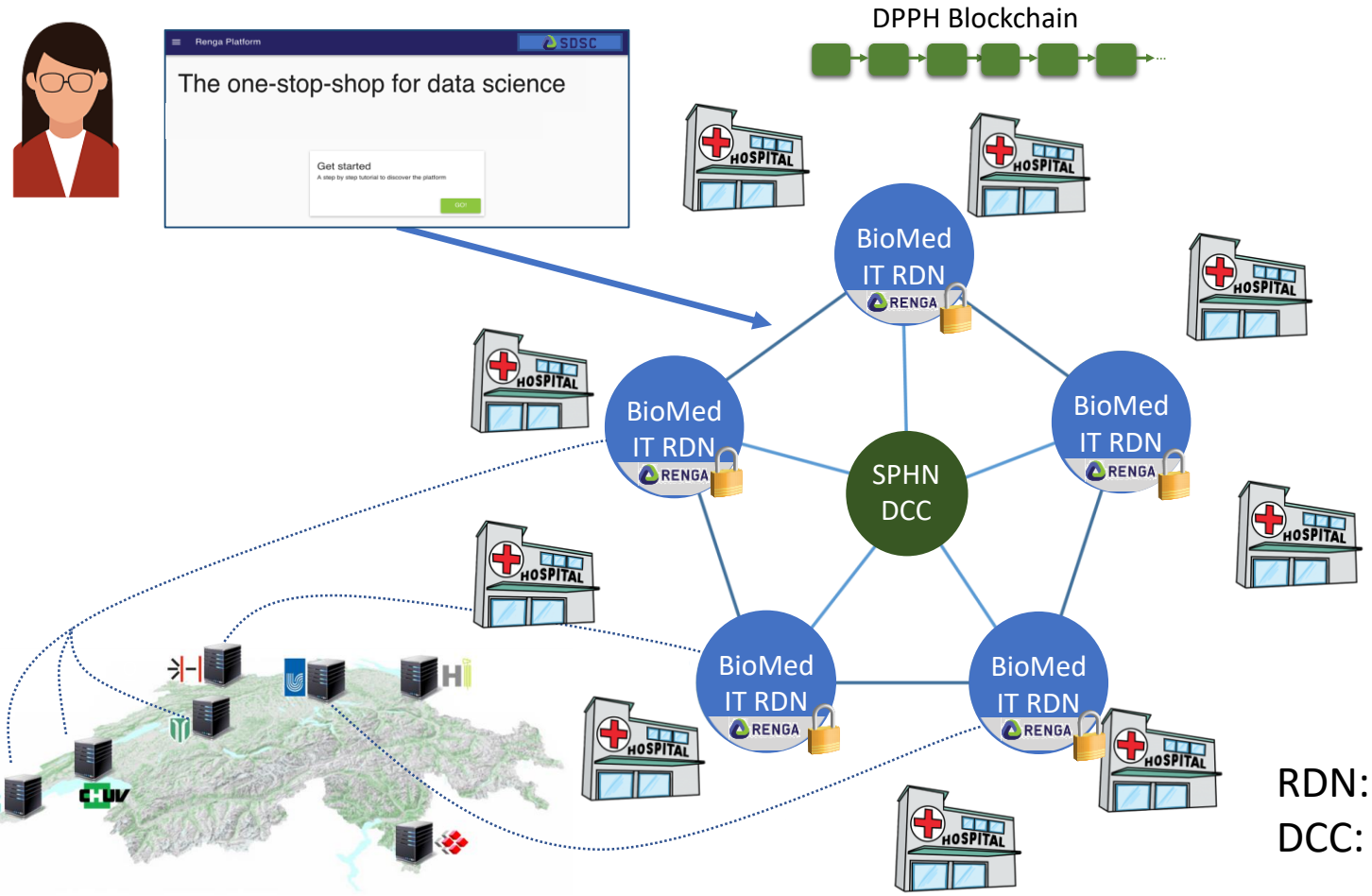
Strategic Focus Area
Personalized Health and Related Technologies

Project goals:

- Address the main **privacy, security, scalability, and ethical challenges** of data sharing for enabling effective P4 medicine
- Define an optimal **balance between usability, scalability and data protection**
- Deploy an appropriate set of **computing tools**

DPPH Long-Term Vision

A One-Stop Shop for Collaborative Research on Health Data in the Context of Swiss Personalized Health Network

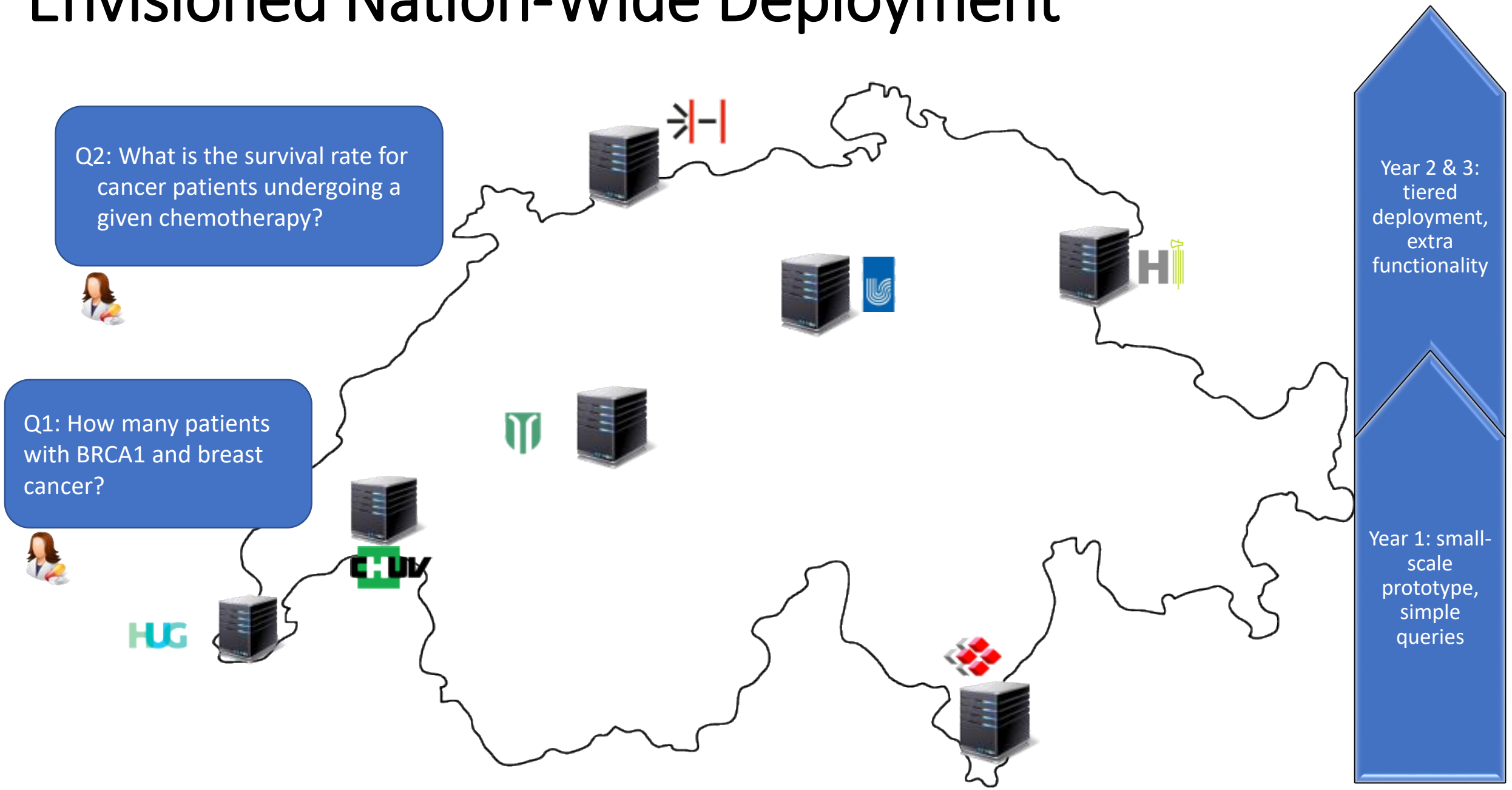


Platform requirements

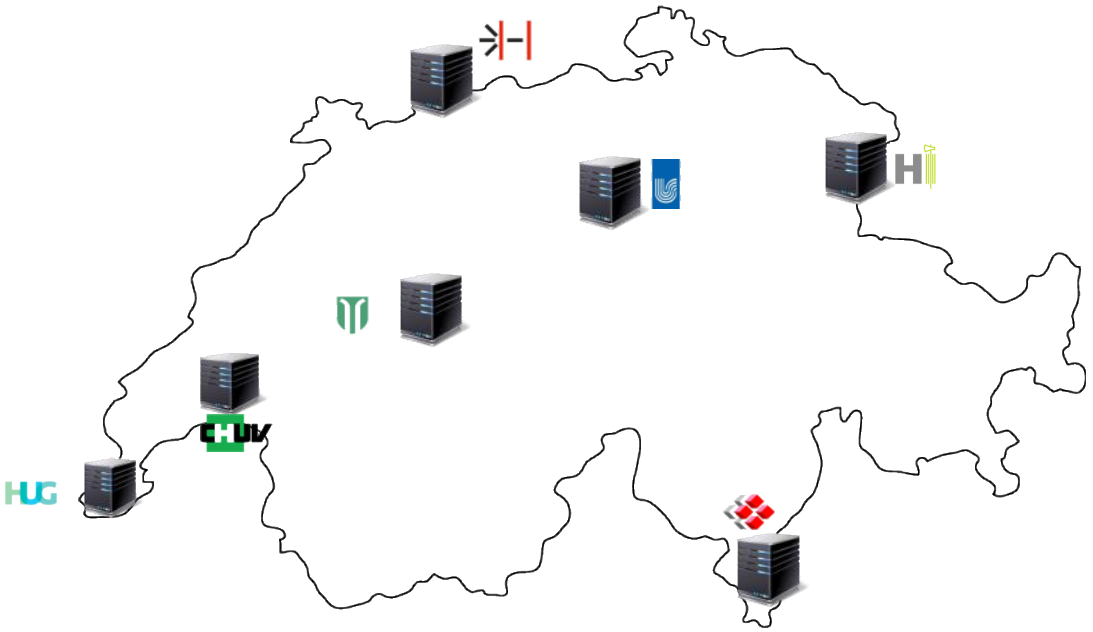
- Interoperability (workflow and data)
- Reproducible research
- Big data scalability
- Auditability and Traceability
- Distributed data
- Secure data access
- Data protection compliance
- Privacy-conscious processing

RDN: Regional Data Node
DCC: Data Coordination Center

Envisioned Nation-Wide Deployment



Data Protection for Personalized Health



Swiss Personalized Health Network

At the international level, see notably:

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

GA4GH has its own workstream on data security

Drynx

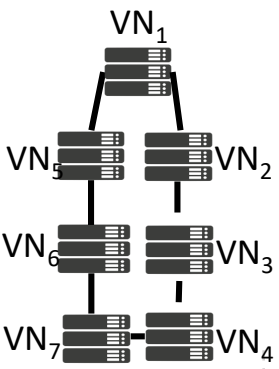
- **Enable computations on federated databases while protecting individual's privacy and data confidentiality**
- In Drynx, no entity has to be individually trusted in order to provide both privacy and robustness
- Adversaries can be malicious (and not just honest-but-curious)

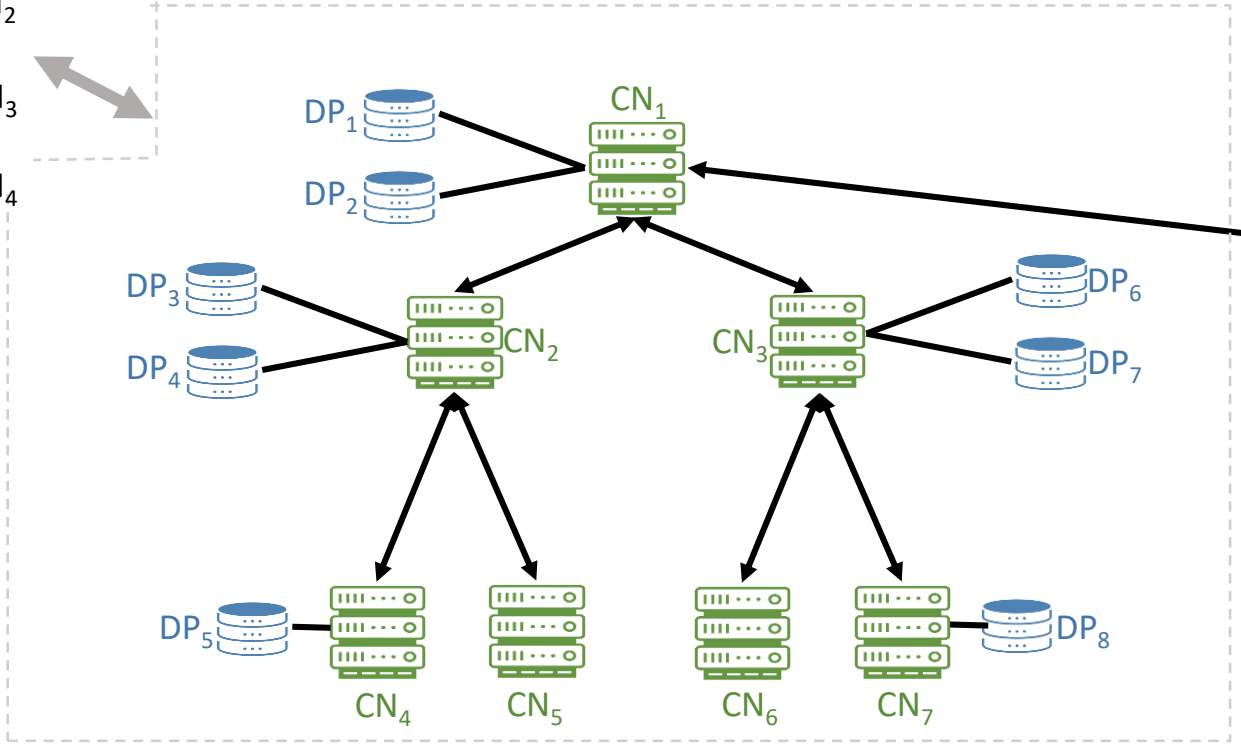
SELECT *sum/count/frequency count*
max/min
variance/standard deviation
and/or
Set intersection/union
Cosine similarity

linear regression
logistic regression

… FROM $DP_1, \ldots, DP_n$ WHERE … GROUP BY/LIKE/… …

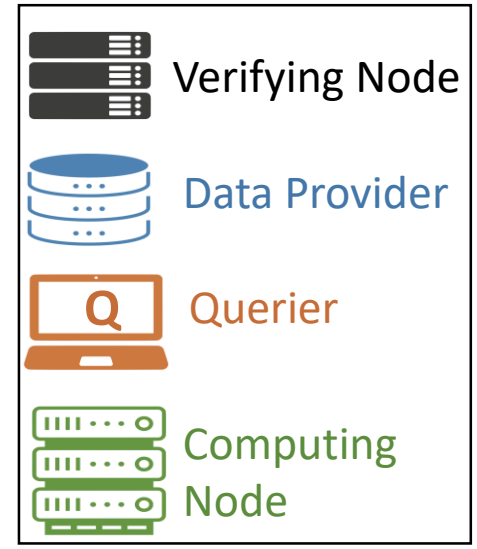
Linear Regression Query (1/5)



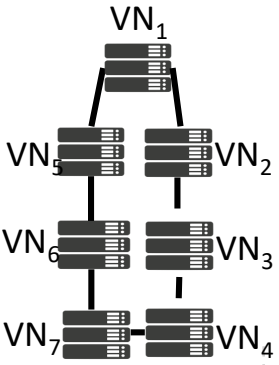


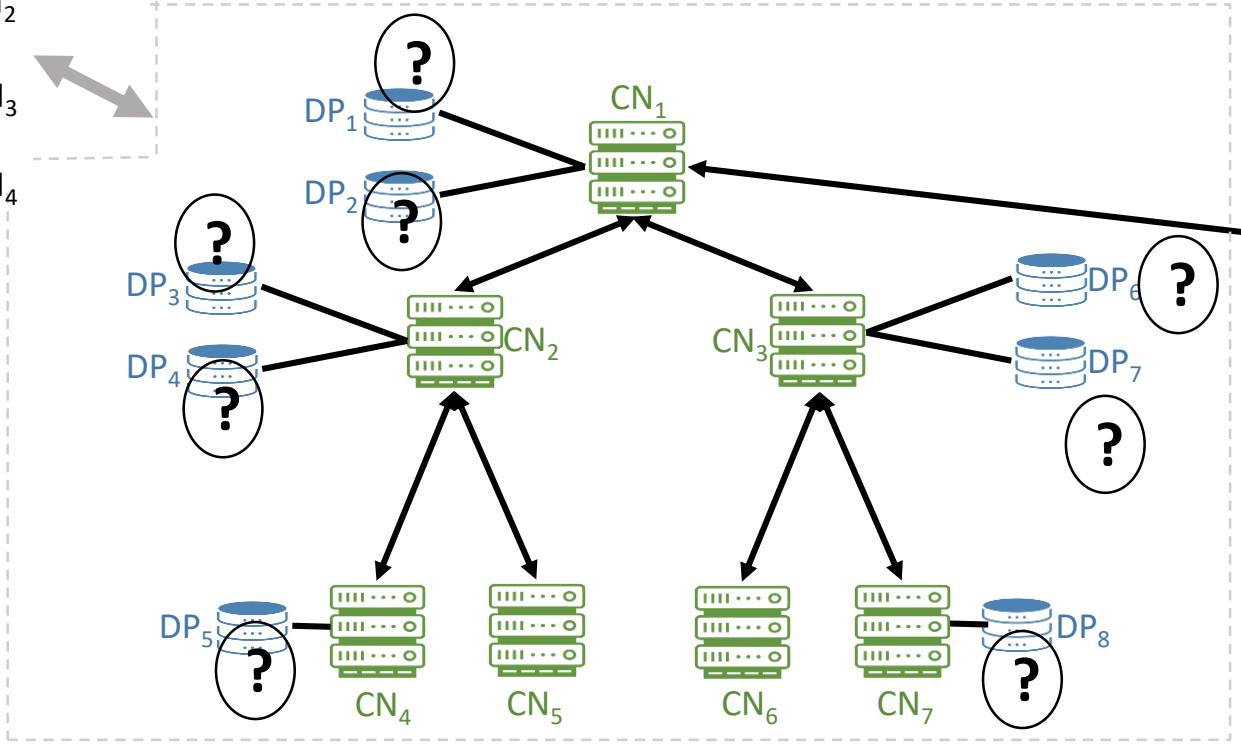
?

SELECT **Linear Regression**
X=[blood_pressure, heartbeat, ...], Y=[cholesterol_rate]
WHERE *patient_age in [50-90]*

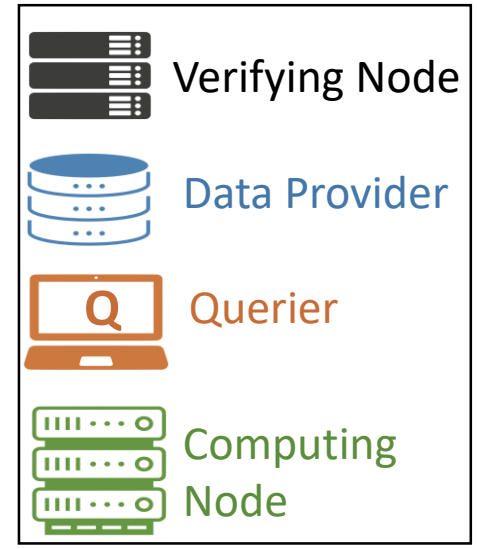


Linear Regression Query (2/5)



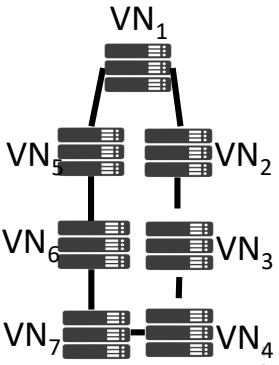


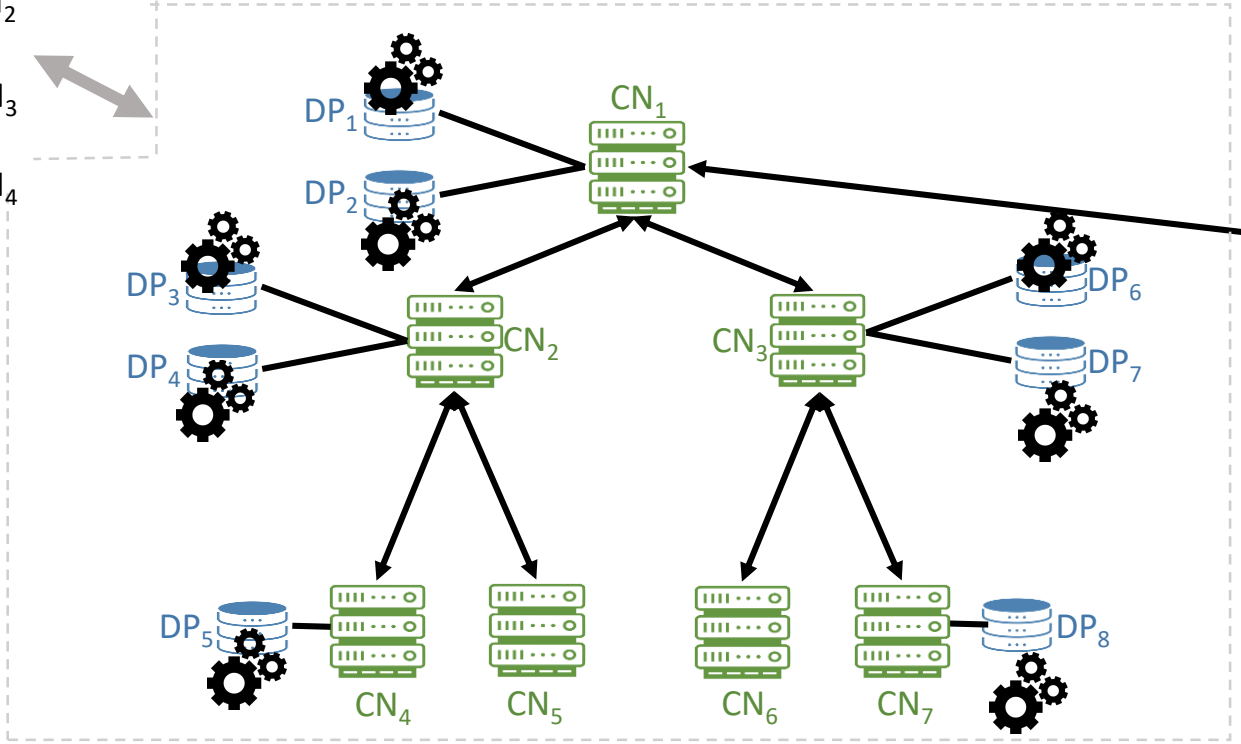
SELECT Linear Regression
X=[blood_pressure, heartbeat, …], Y=[cholesterol_rate]
WHERE *patient_age in [50-90]*

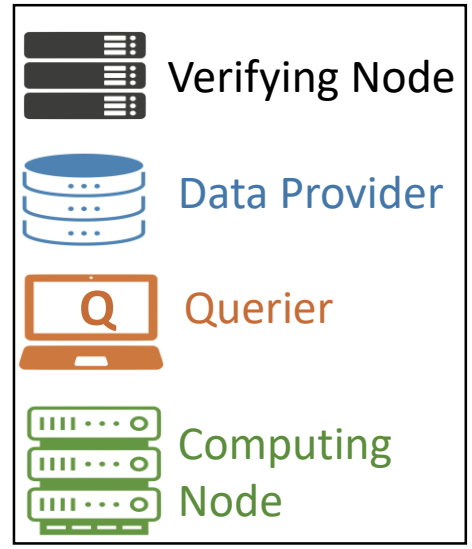


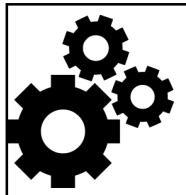
Assume dataset distributed over the DPs with *D* features x_1, …, x_D an a label value *y* such that $y \approx c_0 + c_1x_1 + \ldots + c_Dx_D$

Linear Regression Query (3/5)









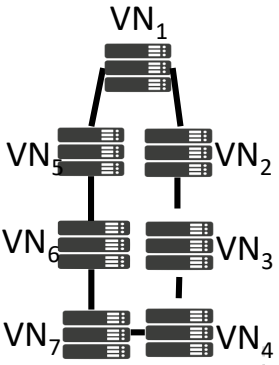
Each DP computes and sends:

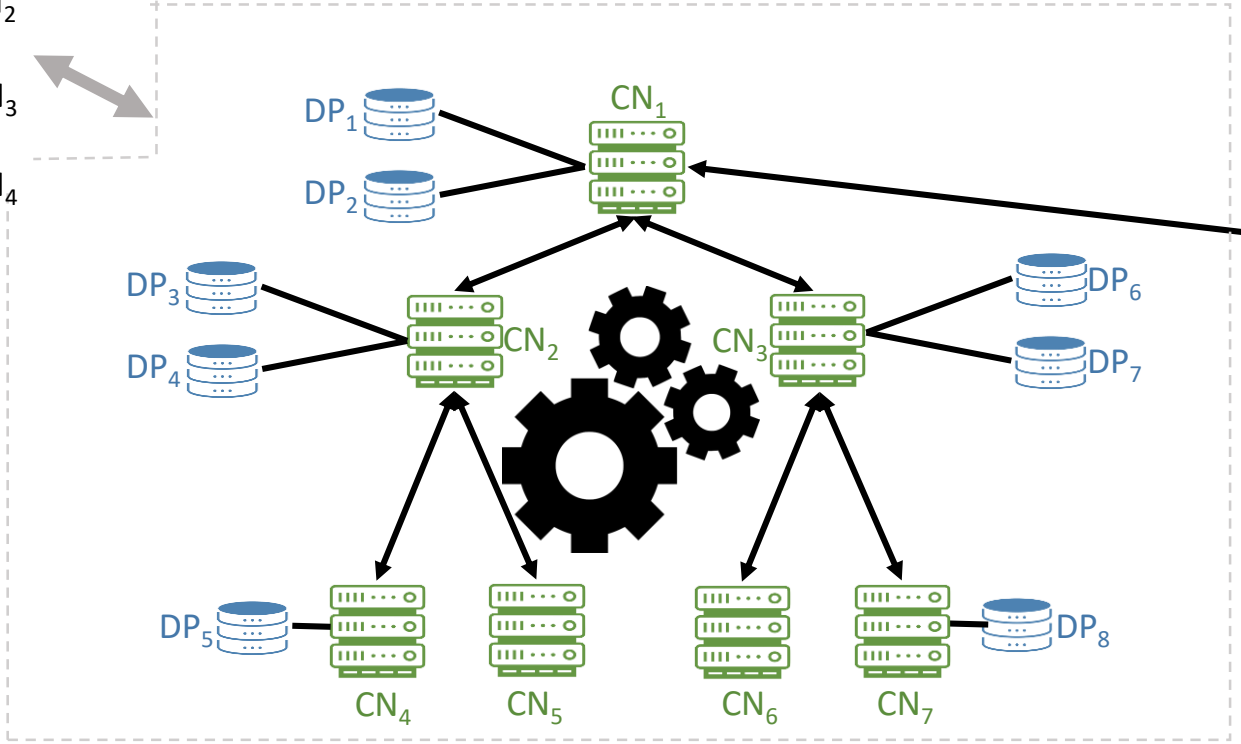
$$\sum_j x_{j,n}\,,\ \sum_j x_{j,n}\,x_{j,c}\,,\ \sum_j y_j\,,\ \sum_j y_j\,x_{j,n}\ d,$$

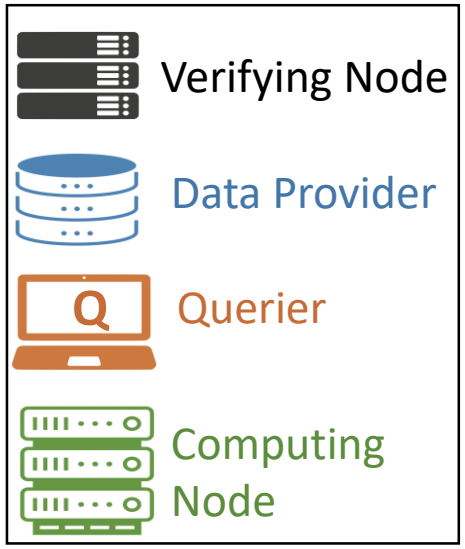
$$\forall\ j,n \in \{1,2,\ldots D\},\ j \neq n$$

Assume dataset distributed over the DPs with *D* features $x_1, \ldots, x_D$ an a label value *y* such that $y \approx c_0 + c_1x_1 + \ldots + c_Dx_D$

Linear Regression Query (4/5)



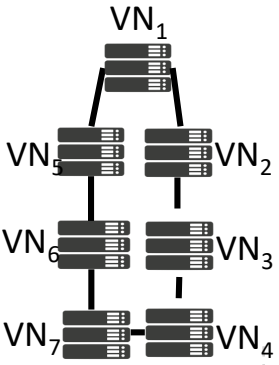


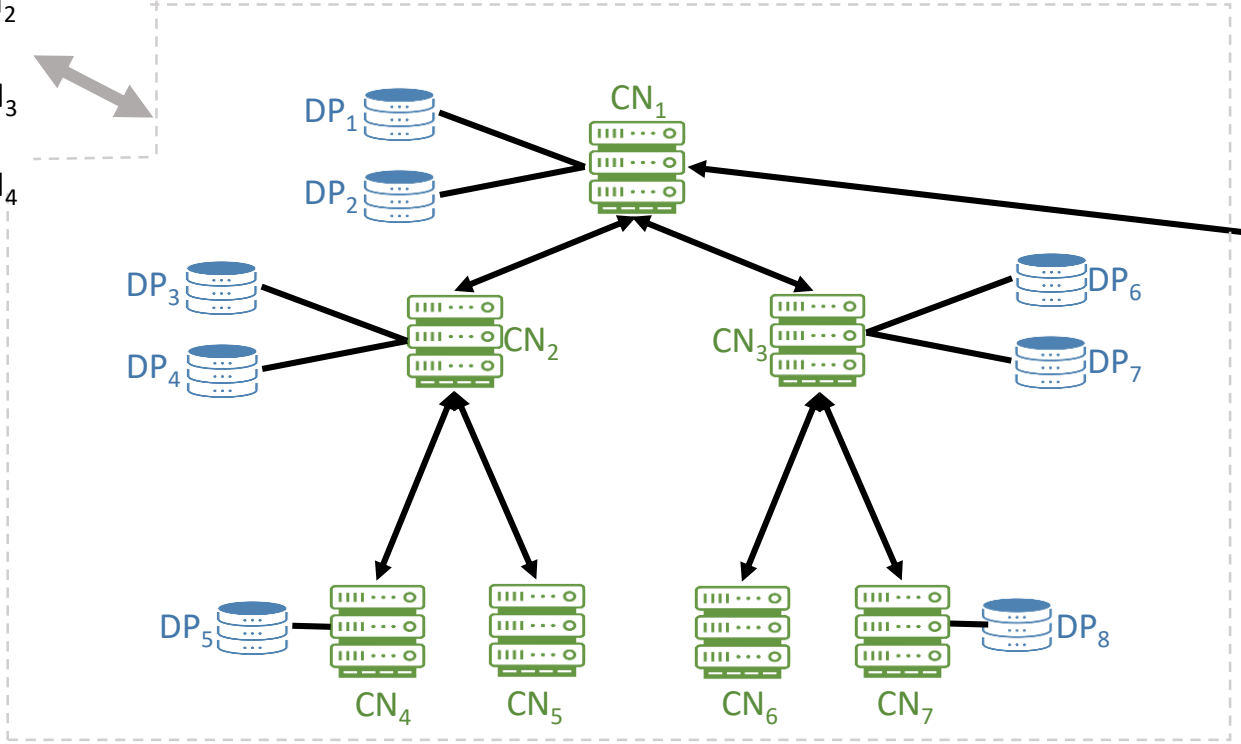


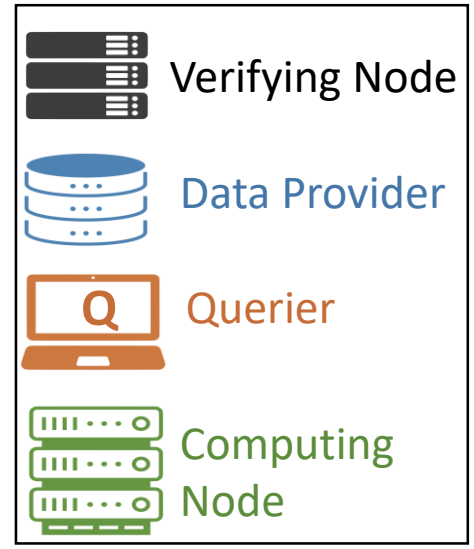
These data are summed by the CNs.

Assume dataset distributed over the DPs with *D* features $x_1, \ldots, x_D$ an a label value *y* such that $y \approx c_0 + c_1x_1 + \ldots + c_Dx_D$

Linear Regression Query (5/5)







Q Such that the querier can solve

$$\begin{bmatrix} n & \sum x_{u,1} & \dots & \sum x_{u,D} \\ \sum x_{u,1} & \sum x^2_{u,1} & \dots & \sum x_{u,1}\, x_{u,D} \\ \dots & \dots & \dots & \dots \\ \sum x_{u,D} & \sum x_{u,1}\, x_{u,D} & \dots & \sum x^2_{u,D} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \dots \\ c_D \end{bmatrix} \approx \begin{bmatrix} \sum y_u \\ \sum y_u\, x_{u,1} \\ \dots \\ \sum y_u\, x_{u,D} \end{bmatrix}$$

All sums are btw. *u= 1* and *u= total nbr. records*

Assume dataset distributed over the DPs with *D* features $x_1, \dots, x_D$ an a label value *y* such that $y \approx c_0 + c_1x_1 + \dots + c_Dx_D$

Drynx also supports Logistic Regression

Relies on an approximation of the logistic regression cost function[1] and achieves an accuracy similar to state of the art privacy solutions.

Data.		Accuracy
LBW	orig.	69.31%
	Drynx	70.26%
PCS	orig.	74.60%
	Drynx	75.13%
Pima	orig.	80.5%
	Drynx	77.55%
SPECTF	orig.	78.9%
	Drynx	74.87%

Scaling factor 10^2; learning rate 0.1; k = 2; l2-regularization factor = 1; 80% training, 20% testing

[1] Y. Aono, T. Hayashi, L. Trieu Phong, and L. Wang. Scalable and Secure Logistic Regression via Homomorphic Encryption. ACM CODASPY, 2016.

LBW = Low birth weight dataset. 9 features http://course1.winona.edu/bdeppa/Biostatistics/Data%20Sets/lowbirtharc.txt

PCS = Prostate Cancer Study. 11 features http://course1.winona.edu/bdeppa/Biostatistics/Data%20Sets/Prostate%20Logistic.txt

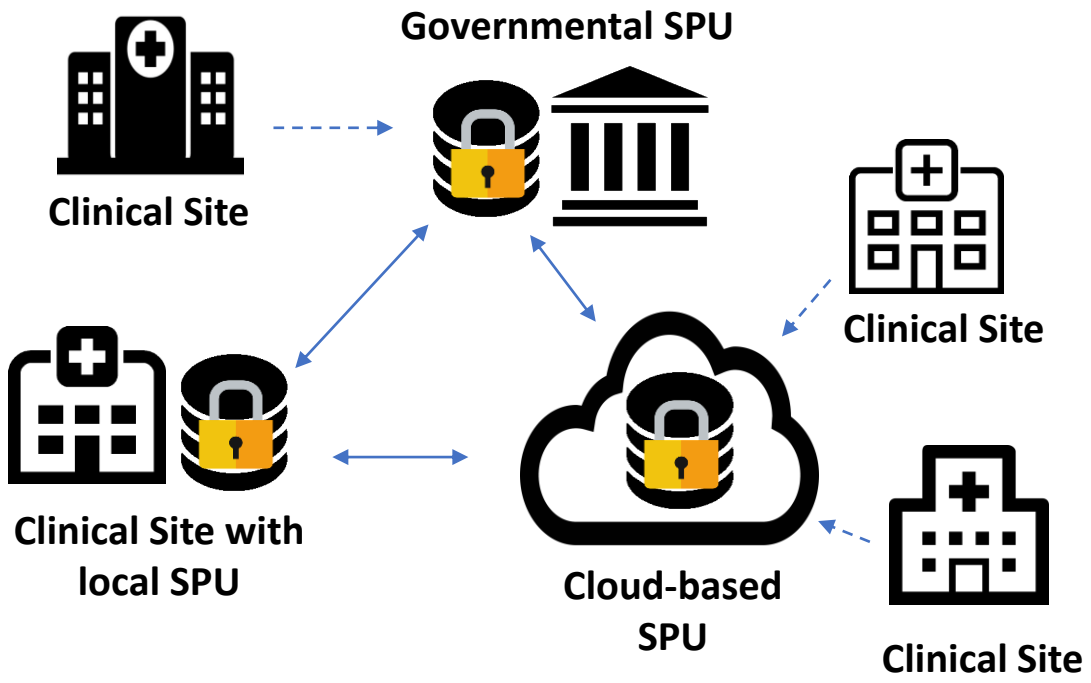
Pima = Pima Indians Diabetes 8 features https://www.kaggle.com/uciml/pima-indians-diabetes-database

SPECTF = Single Proton Emission Computed Tomography. 44 features https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart

Next Step for Drynx: Neural Networks

- As usual in Drynx:
  - Fully decentralized architecture
  - Data stay with each data provider
  - Resistance against colluding, malicious adversaries
- More on this in a few months ☺

MedCo: Customized Drynx for Health



Main features:

- **Secure outsourcing** enabled by **collective encryption** of the data
- **End-to-end data protection** through **homomorphic encryption**
- Compliance wrt to **regulations** (e.g., GDPR)
- **Increased flexibility and lower costs** wrt standard approaches of data sharing

Raisaro JL, Troncoso-Pastoriza JR, Misbach M, Gomes de Sá ES, André J, Pradervand S, Missiaglia E, Michielin O, Ford BA, Hubaux JP. **MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data**. Accepted for publication in IEEE/ACM Transactions in Computational Biology and Bioinformatics

MedCo: Combining the best of Information Security and Medical Informatics

Data model

Interoperability layer
Meta API

Privacy-preserving computing framework

Modern GUI

DISCLAIMER

MedCo is a generic concept and it is not fundamentally tied to these technologies, but can be adapted and integrated to other ones

Events on Genome Privacy and Security

- **Dagstuhl** seminars on genome privacy and security 2013, 2015
- **Conference on Genome and Patient Privacy (GaPP)**
  - March 2016, Stanford School of Medicine
- **GenoPri**: International Workshop on Genome Privacy and Security
  - July 2014: Amsterdam (co-located with PETS)
  - May 2015: San Jose (co-located with IEEE S&P)
  - November 12, 2016: Chicago (co-located with AMIA)
  - October 15, 2017: Orlando (co-located with Am. Society for Human Genetics (ASHG) and GA4GH)
  - October 3, 2018, Basel (co-located with GA4GH)

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

- **iDash**: integrating Data for Analysis, Anonymization and sHaring (already in previous years)
  - **October 14, 2017: Orlando**
- Inst. For Pure and Applied Mathematics (IPAM, UCLA)

  Algorithmic Challenges in Protecting Privacy for Biomed Data

  10-12 January, 2018
- DPPH Workshop, 15 February 2018

iDASH

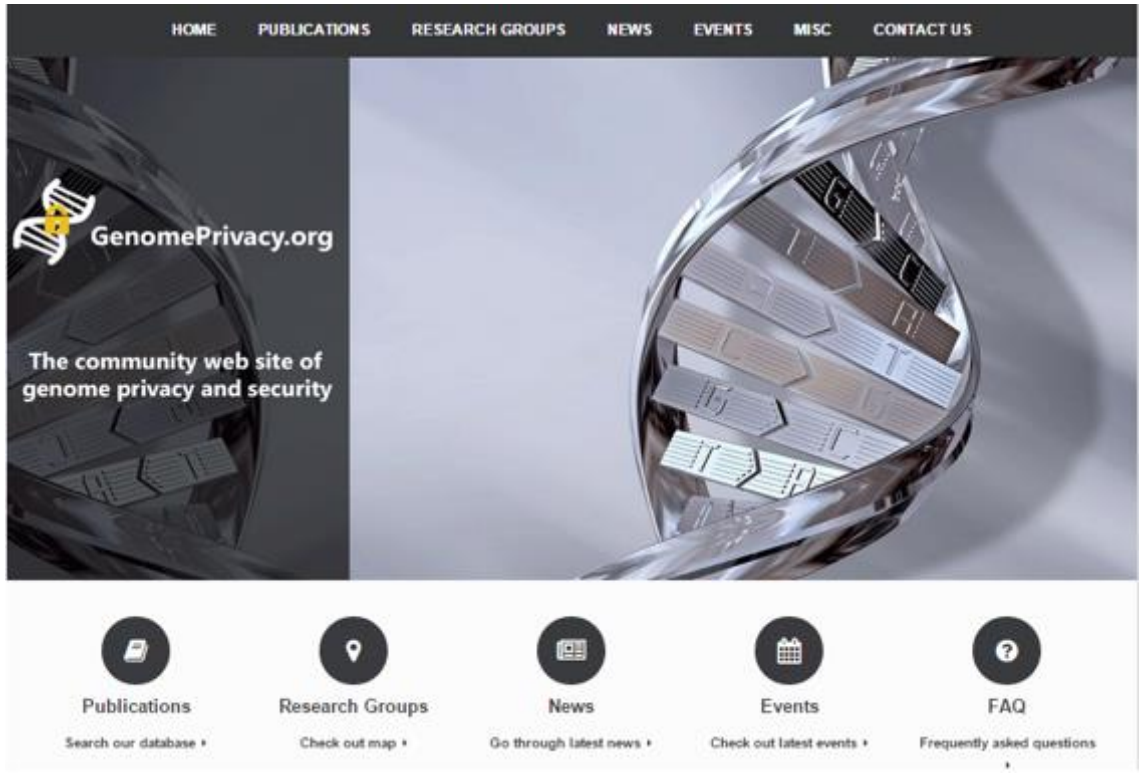
ipam

DPPH18

➔ Lots of material online

DPPH18.epfl.ch

“genomeprivacy.org”



Community website

- Searchable list of publications on genome privacy and security
- News from major media (from Science, Nature, GenomeWeb, etc.)
- Research groups and companies involved
- Tutorial and tools
- Events (past & future)

The Center for Digital Trust @EPFL https://c4dt.org/

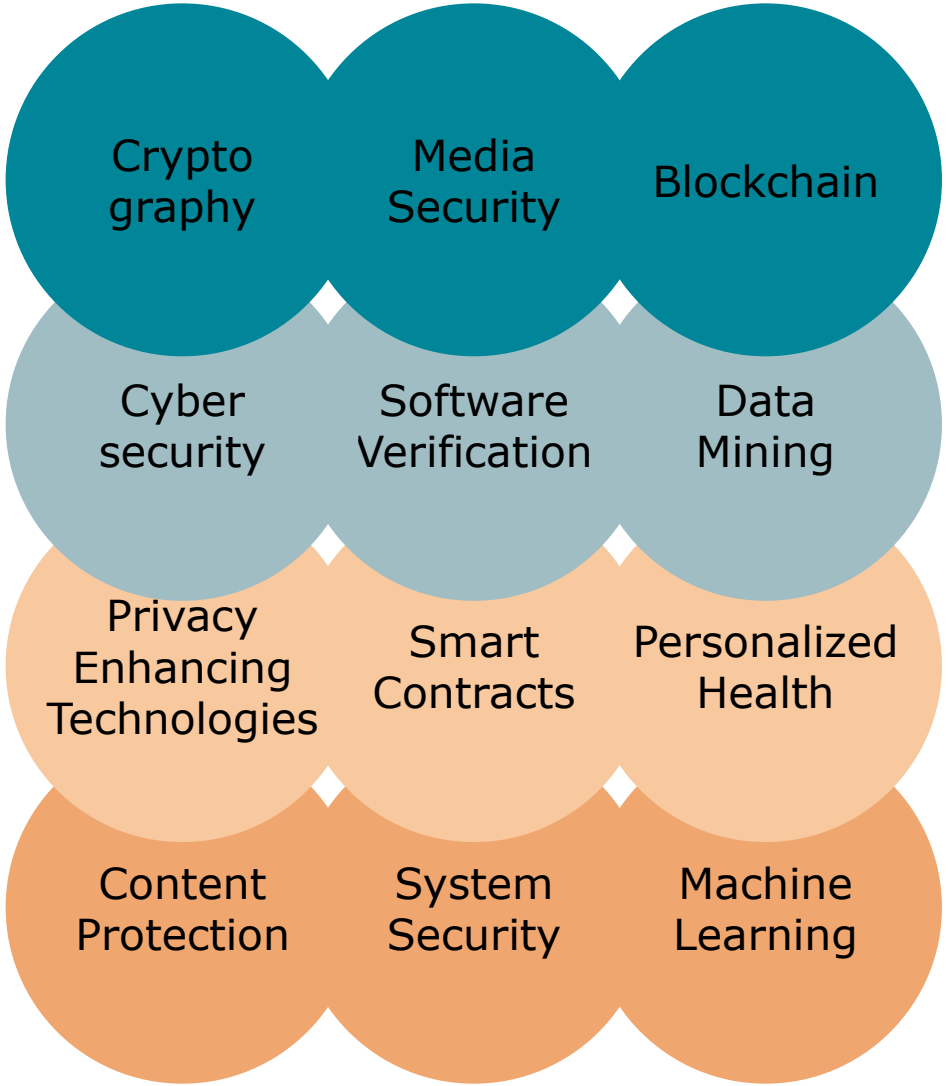
Reinvent trust for the digital society

- Center of expertise
- One-stop-shop
- Community

Collaboration ecosystem

- 30+ EPFL laboratories
- 10+ Organizations

CICR, CHUV, BYZGEN, SWISS POST, Swissquote, ELCA, SICPA, 17 96 LOMBARD ODIER, swisscom, CISCO, Swiss Re, NAGRA KUDELSKI GROUP



Conclusion

- Worldwide, the confidentiality of health data is **in jeopardy**
- Standardization and regulation of AI in health can only be achieved if people trust the whole process to be **safe, secure** and **fair**
- Advanced privacy-enhancing technologies can be **effective enablers** to achieve FG-AI4H goals
- Synergy to be explored between FG-AI4H and GA4GH (Global Alliance for Genomics and Health)
- Check our projects and tools:
  - https://DPPH.ch
  - https://MedCo.epfl.ch