Inselspital
Bern Seminar Series for Precision Medicine

15 October 2019

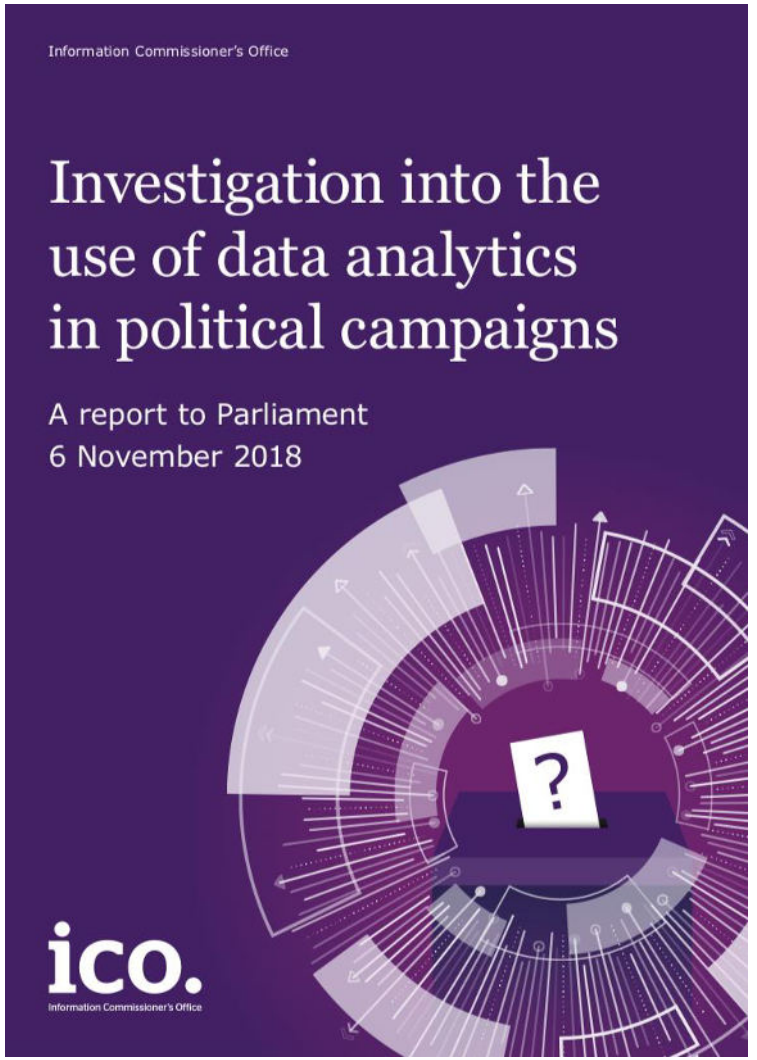
Data Protection for Personalized Health

Prof. Jean-Pierre Hubaux
Head of the Laboratory for Data Security
Academic Director of the Center for Digital Trust
School of Computer and Communication Sciences
EPFL

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

2016: Massive voter manipulation

“Brexit vote” and US presidential elections
→ Two major democracies find themselves internally polarized, victim of home-made digital tools



Information Commissioner’s Office (UK's independent body set up to uphold information rights)

Will Democracy Survive Big Data Breaches?

Cambridge Analytica had around 5000 data points on each targeted voter, provided by Facebook.

What if it had access to more?

“There is always going to be a Cambridge Analytica”

US Healthcare Official "Wall of Shame"

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

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

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:

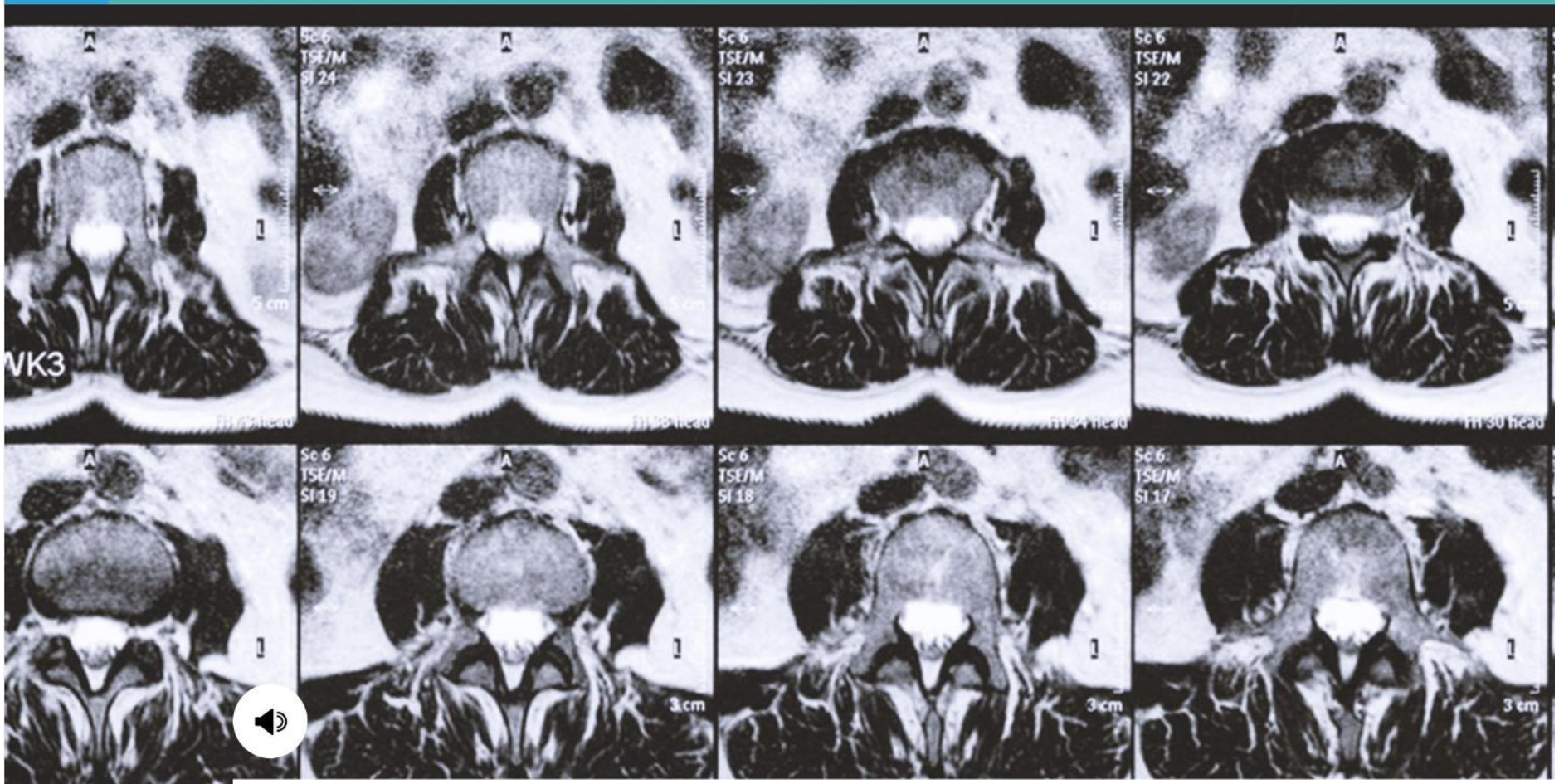
Cases Currently Under Investigation

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



17.09.2019, 07:02 Uhr

Millionenfach Patientendaten ungeschützt im Netz

Hochsensible medizinische Daten, unter anderem von Patienten aus Deutschland und den USA, sind nach Recherchen des BR und der US-Investigativplattform ProPublica auf ungesicherten Internetservern gelandet. Jeder hätte darauf zugreifen können.

ProPublica ProPublica Illinois Local Reporting Network Electionland Donate

PROPUBLICA Graphics & Data Newsletters About Get the Big Stor Join

Trump Administ Immigrat Health C Educat More. Serie Vide Impac Searc

Jason Raish, special to ProPublica

Millions of Americans' Medical Images and Data Are Available on the Internet. Anyone Can Take a Peek.

Hundreds of computer servers worldwide that store patient X-rays and MRIs are so insecure that anyone with a web browser or a few lines of computer code can view patient records. One expert warned about it for years.

by Jack Gillum, Jeff Kao and Jeff Larson, Sept. 17, 12 a.m. EDT

"Legal deterrence" and public shame are clearly not enough!

‹ Gesetzgebung

Gesetzgebung Mensch & Gesundheit

Gesetzgebung Elektronisches Patientendossier (EPDG)

Gesetzgebung Arzneimittel und Medizinprodukte

Gesetzgebung Betäubungsmittel

Gesetzgebung Transplantationsmedizin

Gesetzgebung Genetische Untersuchungen

Gesetzgebung Elektronisches Patientendossier (EPDG)

Das Bundesgesetz über das elektronische Patientendossier regelt die Rahmenbedingungen für die Einführung und Verbreitung des elektronischen Patientendossiers und tritt am 15. April 2017 in Kraft.

Mit dem elektronischen Patientendossier sollen die Qualität der medizinischen Behandlung gestärkt, die Behandlungsprozesse verbessert, die Patientensicherheit erhöht und die Effizienz des Gesundheitssystems gesteigert sowie die Gesundheitskompetenz der Patientinnen und Patienten gefördert werden.

www.patientendossier.ch

www.e-health-suisse.ch

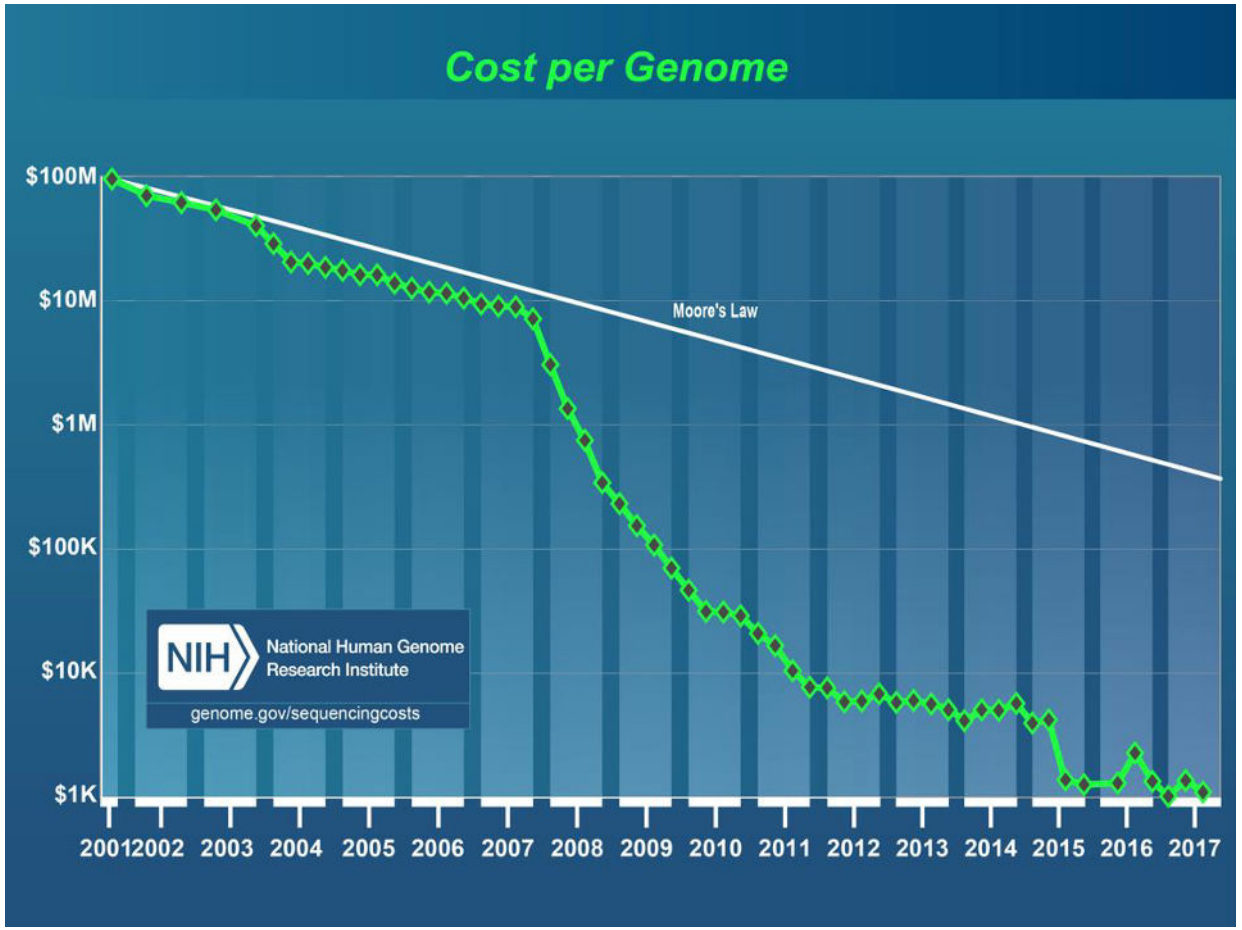
Kontakt

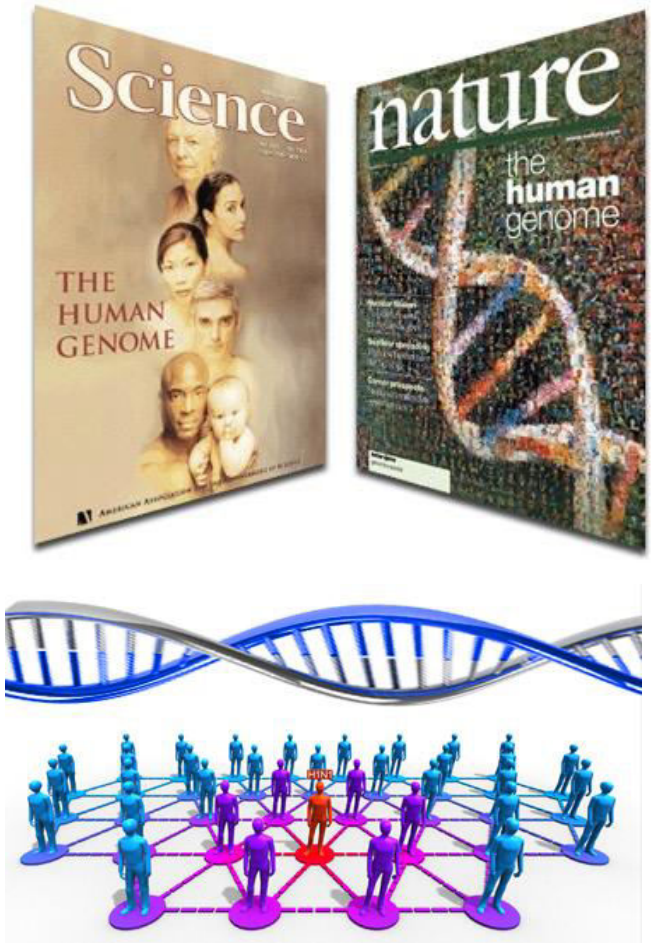
Bundesamt für Gesundheit BAG
Abteilung Gesundheitsstrategien
Sektion eHealth und Krankheitsregister
Schwarzenburgstrasse 157
3003 Bern
Schweiz
Tel. +41 58 462 74 17
E-Mail

Kontaktinformationen drucken

The Genomic Avalanche Is Coming…

Personalized Health





The massive digitalization of clinical and genomic information is providing unprecedented opportunities for improvements in diagnosis, preventive medicine and targeted therapies

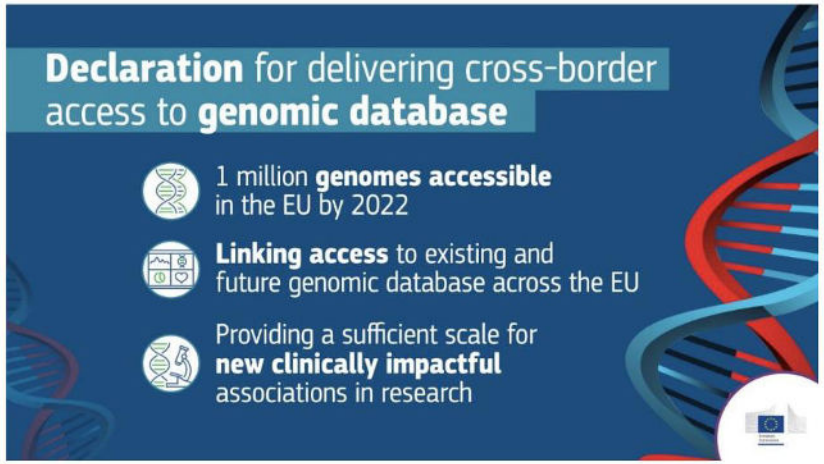
European Commission > Strategy > Digital Single Market > Policies >

Digital Single Market

POLICY

European '1+ Million Genomes' Initiative

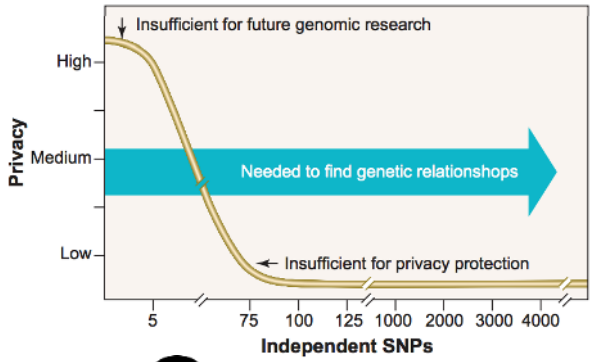
The Signatories of the declaration of cooperation “Towards access to at least 1 million sequenced genomes in the EU by 2022” are setting up a collaboration mechanism with the potential to improve disease prevention, allow for more personalised treatments and provide a sufficient scale for new clinically impactful research.



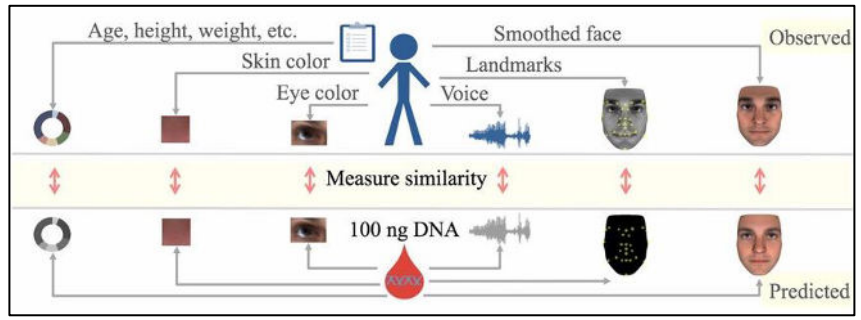
Initiative launched in April 2018

De-identification of genomic data is impossible

- ***Lin et al. 2004 Science***: 75 or more SNPs (Single Nucleotide Polymorphisms) are sufficient to identify a single person
- ***Homer et al. 2008 PLOS Genetics:*** aggregated genomic data (i.e., allele frequencies) can be used for re-identifying an individual in a case group with a certain disease
- ***Gymrek et al. 2013 Science:*** surnames can be recovered from personal genomes, linking "anonymous" genomes and public genetic genealogy databases
- ***Lipper et al. 2017 PNAS:*** Anonymous genomes can also be identified by inferring physical traits and demographic information
- ***Many more to come…***

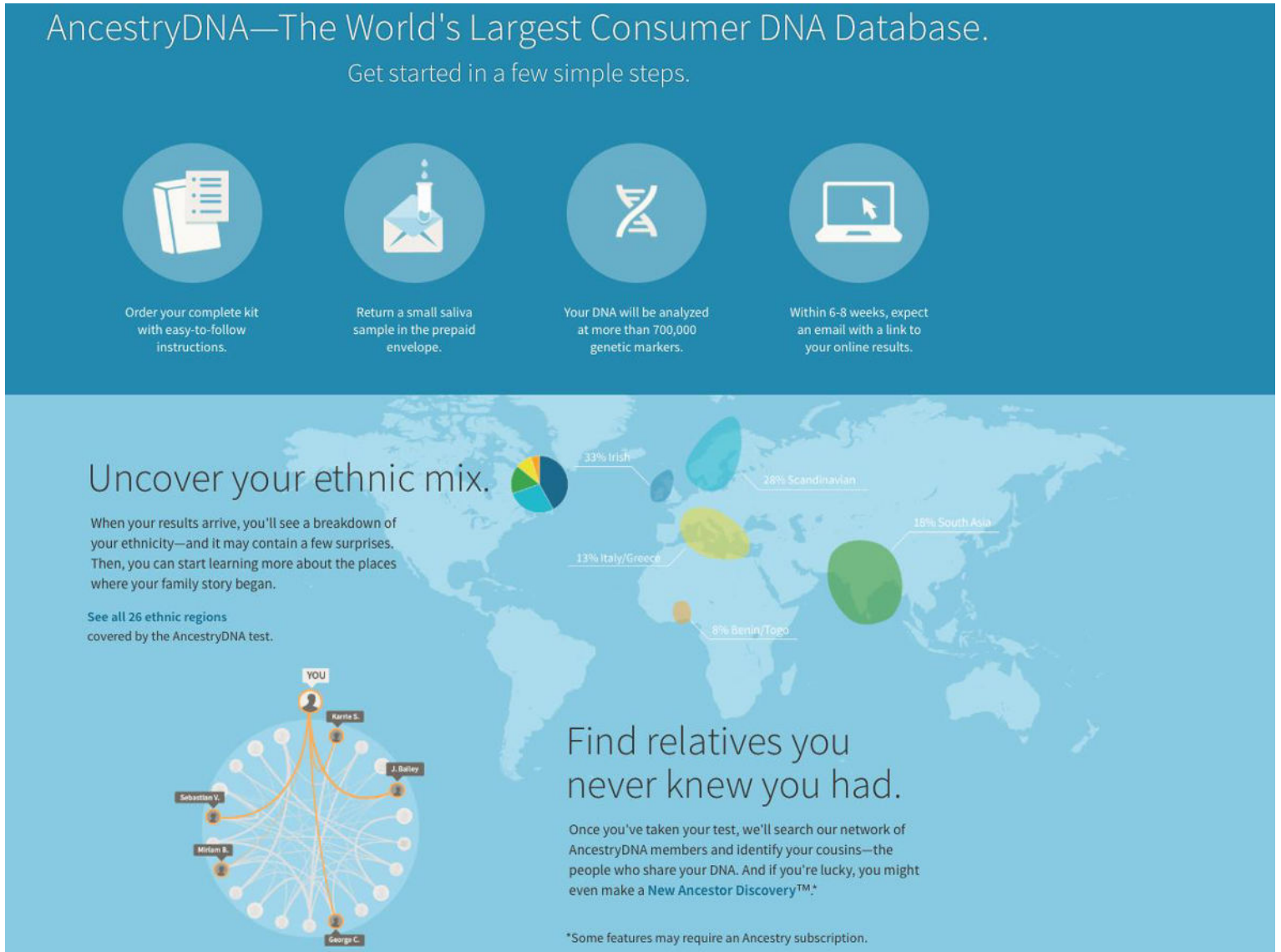


ysearch



Direct-to-Consumer Genomics (1/2)

- Ancestry.com (millions of customers)



Direct-to-Consumer Genomics (2/2)

- 23andMe.com
  (millions customers)

Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson's Disease	★★★★	2.2%	1.6%

Vox GENETICS

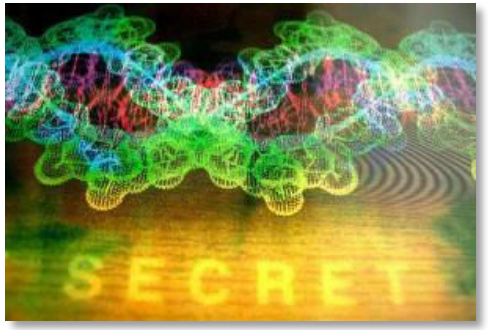
With genetic testing, I gave my parents the gift of divorce

Updated by George Doe on September 9, 2014, 7:50 a.m. ET

TWEET (2,073) SHARE (15K) +

Genome Privacy and Security: a Grand Challenge for Mankind

- Required **duration** of protection >> **1 century**
- (Current) **data size**: around **300 GBytes** / person
- Need sometimes to carry out computations on **millions** (if not more) of patient records
- **Noisy** data
- **Correlations**
  - within a single genome ("linkage disequilibrium")
  - across genomes (kinship, ethnicity)



- **Several "semi-trusted" stakeholders**: sequencing facilities (including Direct-to-Consumer companies), hospitals, genetic analysis labs, private doctors,…
- **Diversity of applications** (and thus of requirements): healthcare, medical research, forensics, ancestry

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

Multi-site Studies – Where to Store the Data?

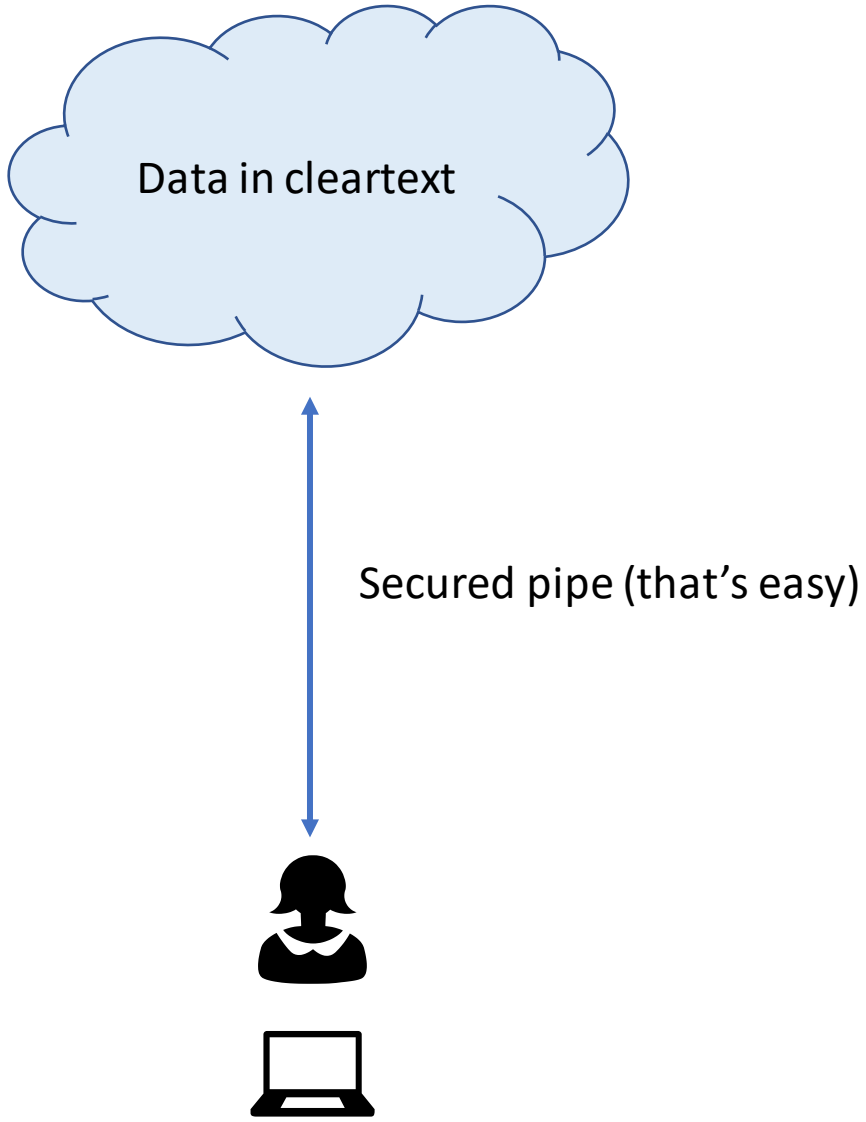
a. Keep them at each site

- Useful especially if the cloud is untrusted
- Better control of the data

b. In the cloud

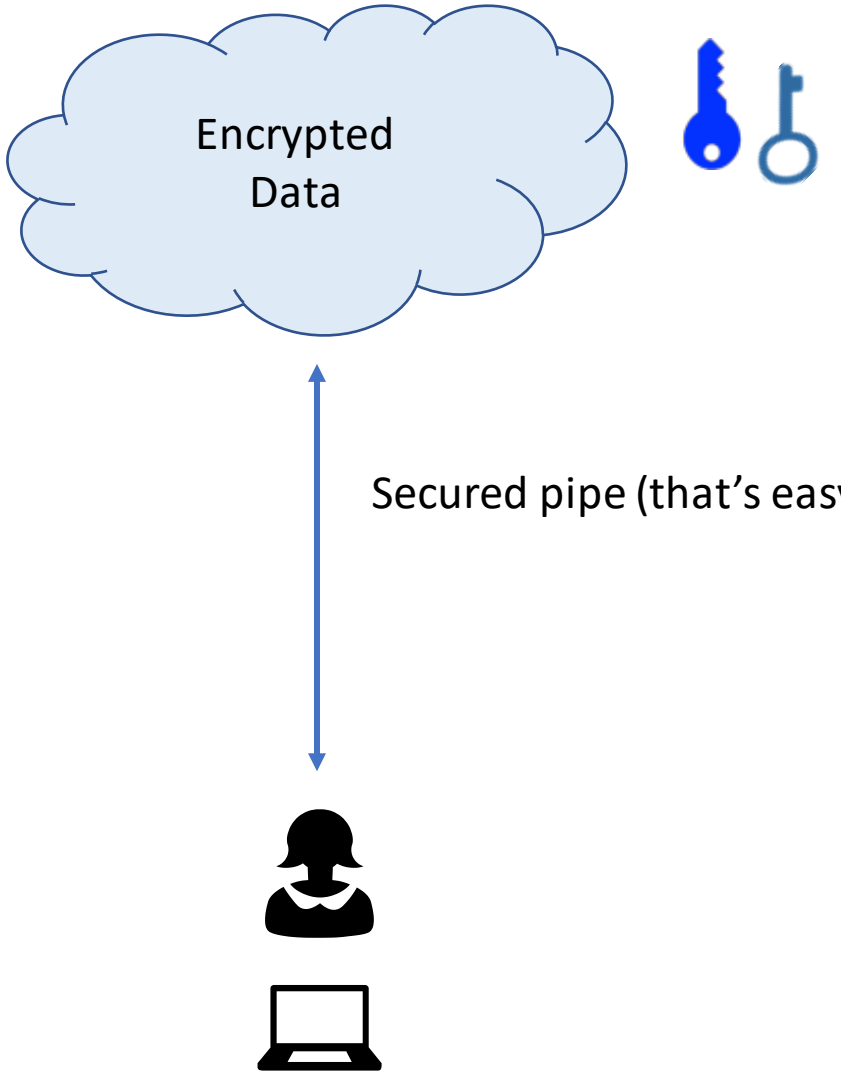
- Take advantage of the well-known strengths of the cloud (see next slide)

Case 0: The Cloud is Fully Trusted – Storage in clear text (never happens in practice)



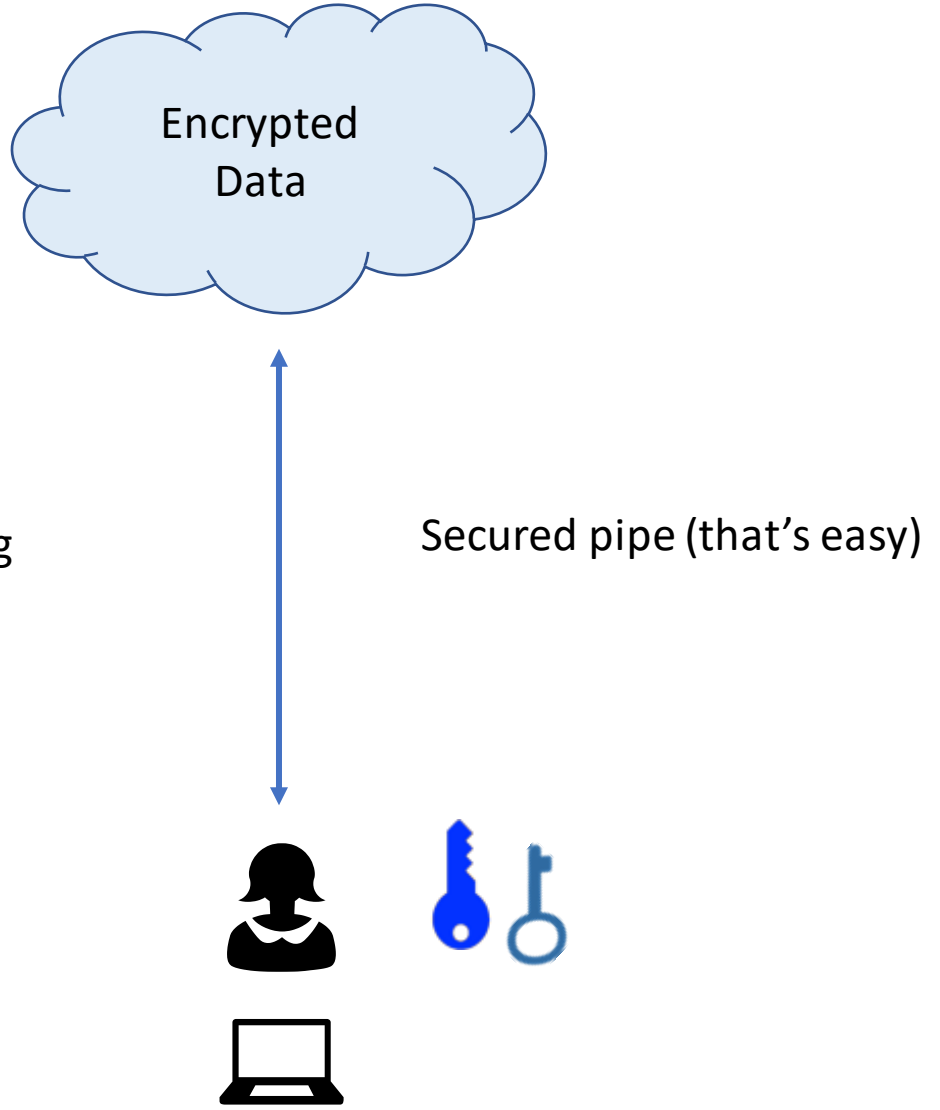
- Data sharing is easy
- Computation in the cloud is easy

Case 1: The Cloud is Fully Trusted – It encrypts with keys it controls



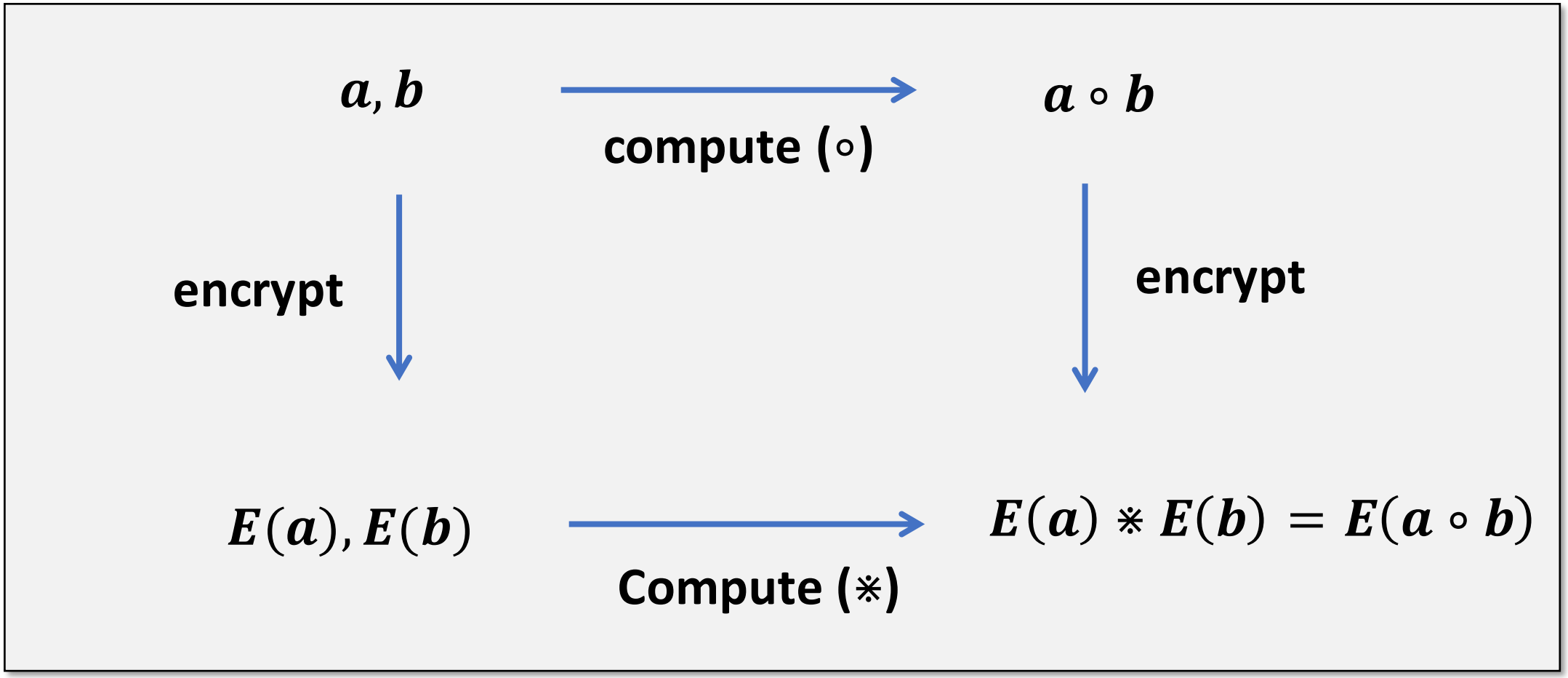
- Data sharing is easy
- Computation in the cloud is easy

Case 2: The Cloud Is Untrusted – The user encrypts under their own keys



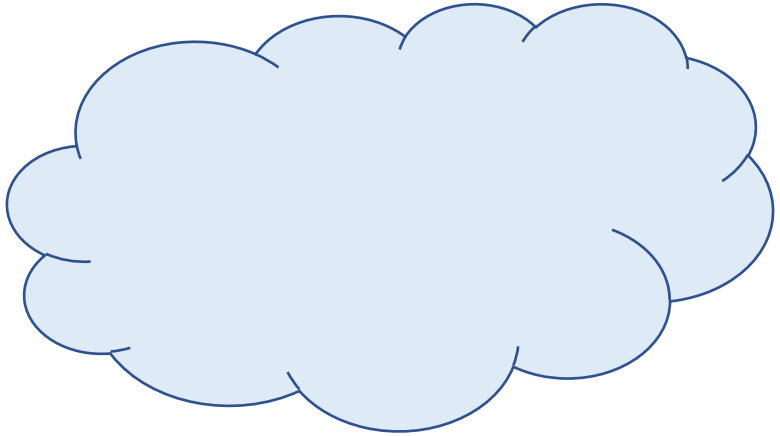
- Data sharing is tricky (key management)
- Computation in the cloud is impossible
- Some of the benefits of cloud computing are thus lost
- If the user loses their keys, they lose all their data

Homomorphic Encryption



Homomorphic encryption enables computations directly on encrypted data.

Case 3: The Cloud is Untrusted – The user homomorphically encrypts with keys it controls (1/3)



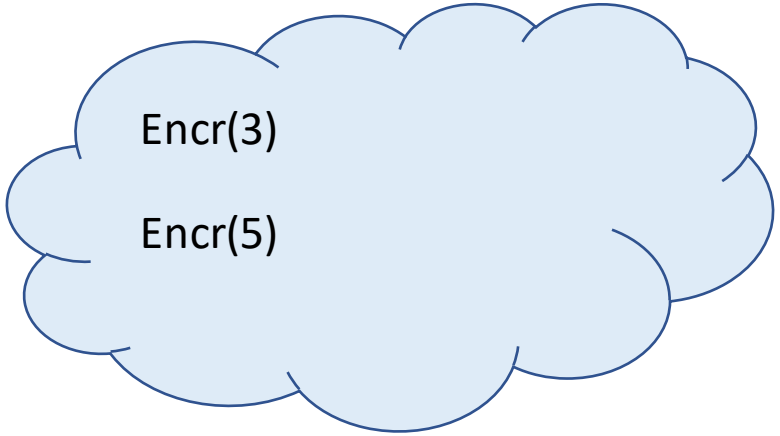
- Data sharing is doable
- Computation in the cloud is possible, but expensive

Secured pipe

Encr(3)

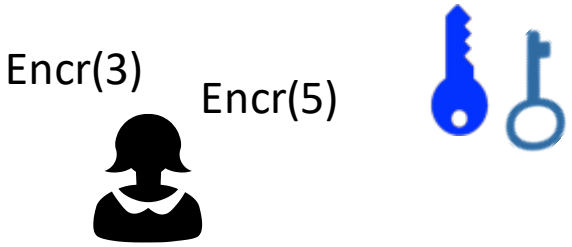
Encr(5)

Case 3: The Cloud is Untrusted – The user homomorphically encrypts with keys it controls (2/3)

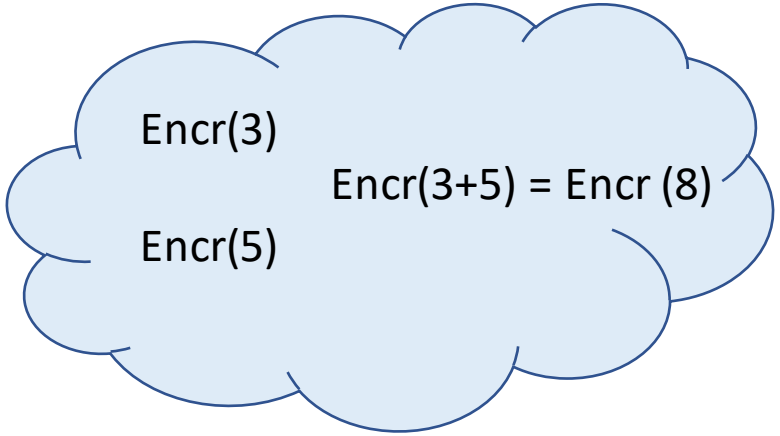


- Data sharing is doable
- Computation in the cloud is possible, but expensive

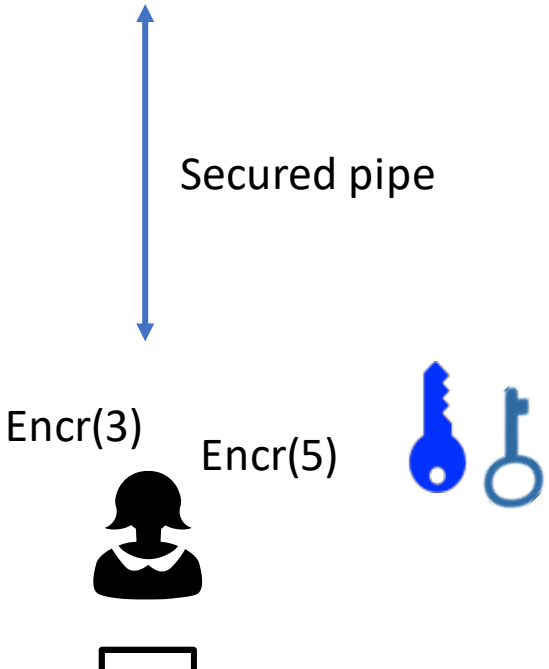
Secured pipe



Case 3: The Cloud is Untrusted – The user homomorphically encrypts with keys it controls (3/3)



- The cloud can make computations on encrypted data, **for which it does not know the crypto keys**
- Hence computation in the cloud is possible (albeit expensive)
- Data sharing is doable



Multi-site Studies: Keeping the Data at Each Site

Assume Sites do not trust each other
→ Possible solution: Secure Multi-Party Computation

Secure Multiparty Computation

Problem statement:

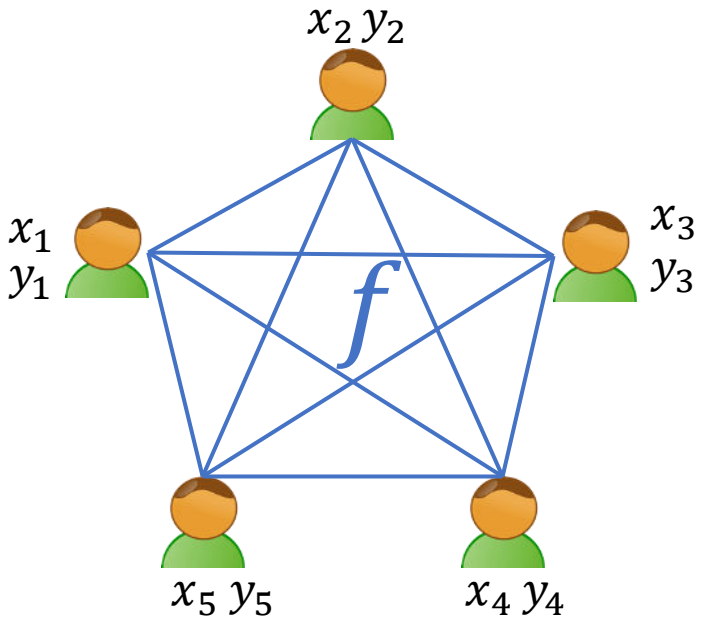
A set of players $\mathcal{P} = \{P_1, P_2, \dots, P_N\}$ would like to compute a function $f(x_1, x_2, \dots, x_N) = (y_1, y_2, \dots, y_N)$ of their joint inputs.

Requirements:

1. *Privacy*
   No party should learn anything more than its prescribed output
2. *Correctness*
   Each party is guaranteed that the output that it receives is correct

Realization:

A multiparty cryptographic protocol



Precision Medicine Programs in Switzerland

SAMW ASSM

SIB Swiss Institute of Bioinformatics

Swiss Personalized Health Network

SPHN

68 MCHF
2017 - 2021

https://www.sphn.ch

ETH zürich

Strategic Focus Area

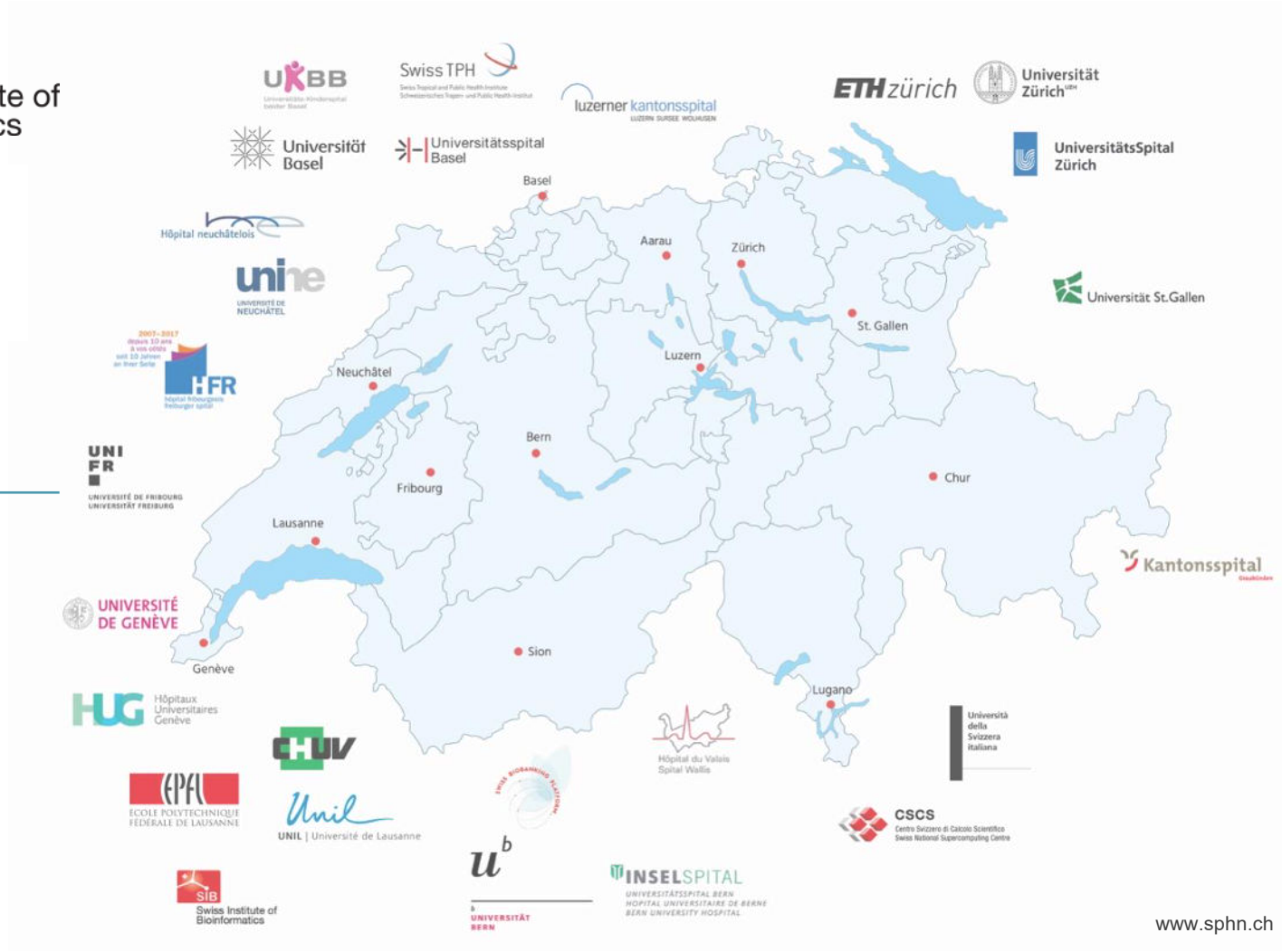
Personalized Health and Related Technologies

50 MCHF
2017 - 2021

https://www.sfa-phrt.ch/

Swiss Biobanking Platform

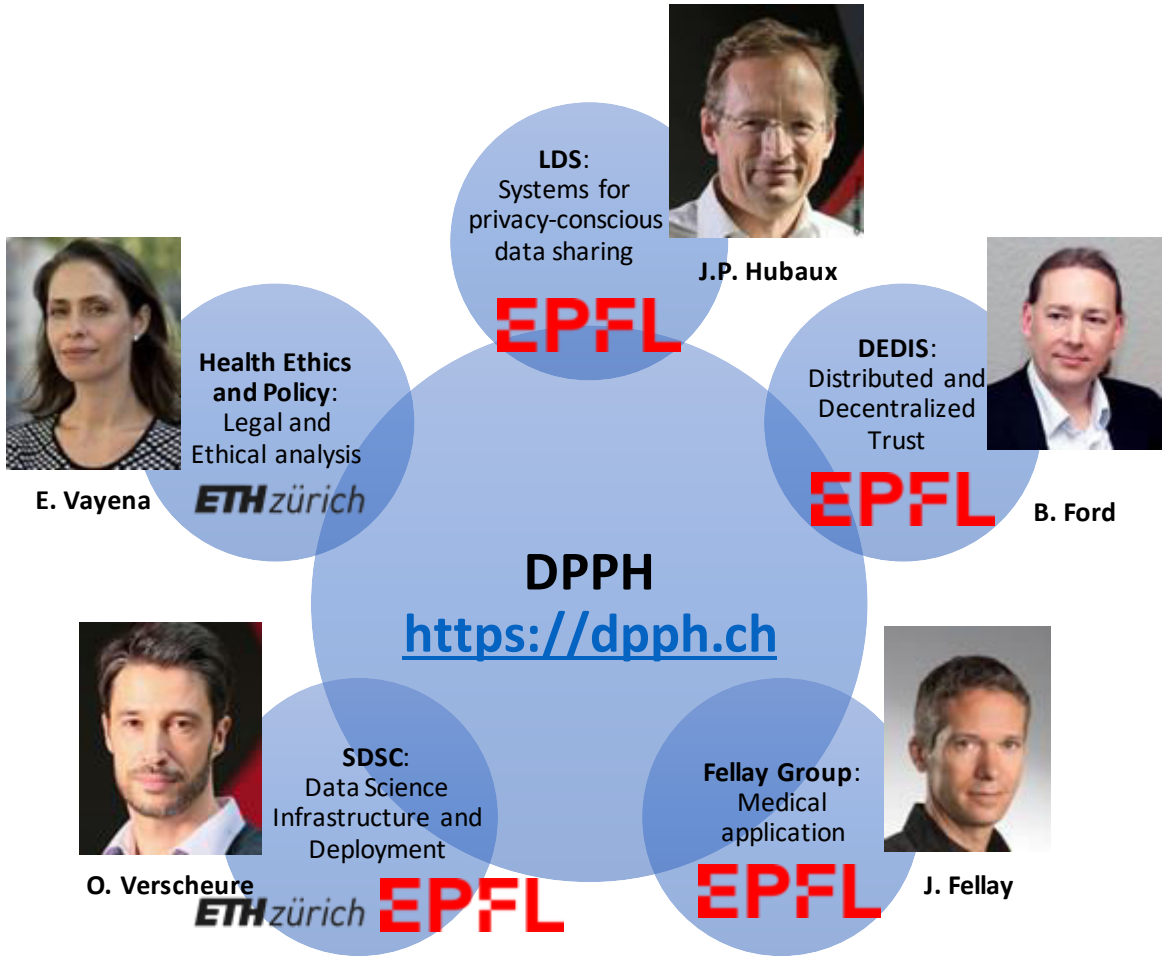
FNSNF
Fonds National Suisse
Schweizerischer Nationalfonds
Fondo Nazionale Svizzero
Swiss National Science Foundation



www.sphn.ch

Data Protection in Personalized Health

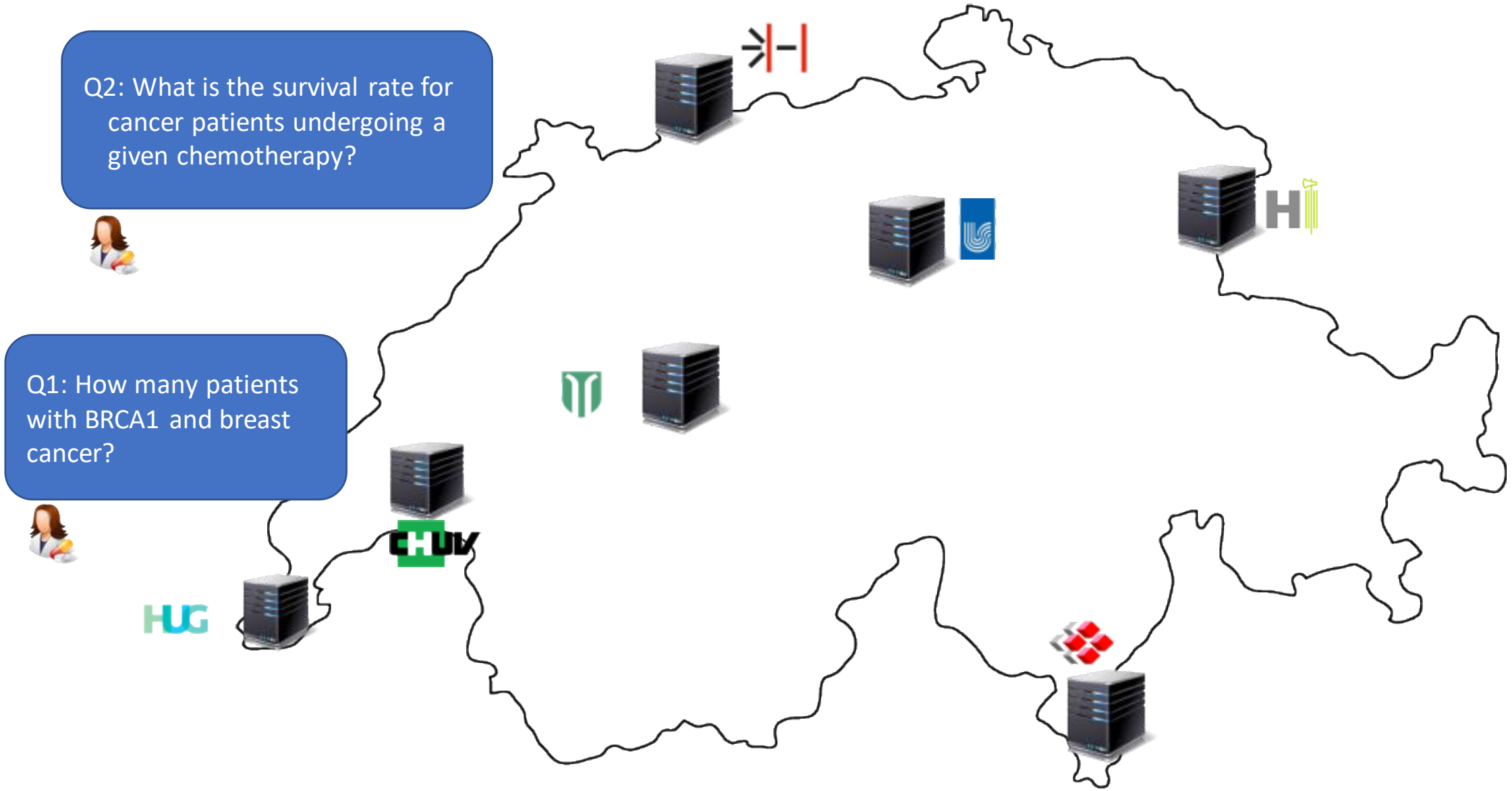
- 4 research groups across the ETH domain + SDSC (Swiss Data Science Center)
- Funding: 3 Millions CHFrs
- Duration: 3 years (4/2018 - 3/2021)
- Funding Program: ETH PHRT (Personalized Health and Related Technologies)



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**

Envisioned Nation-Wide Deployment



MedCo: Consortium and project goals

- Funding: SPHN + PHRT
- Budget: 530K CHF
- Start date: April 1st 2019
- Duration: 18 months
- First application: oncology: O. Michielin,…
- Goal(s):
  1. Bringing MedCo from an “academic” prototype to “hospital-compliant” operational system
  2. Deploy and test MedCo in (at least) 3 Swiss University Hospitals
  3. Validate MedCo with end-users

N. Rosat

J. Fellay

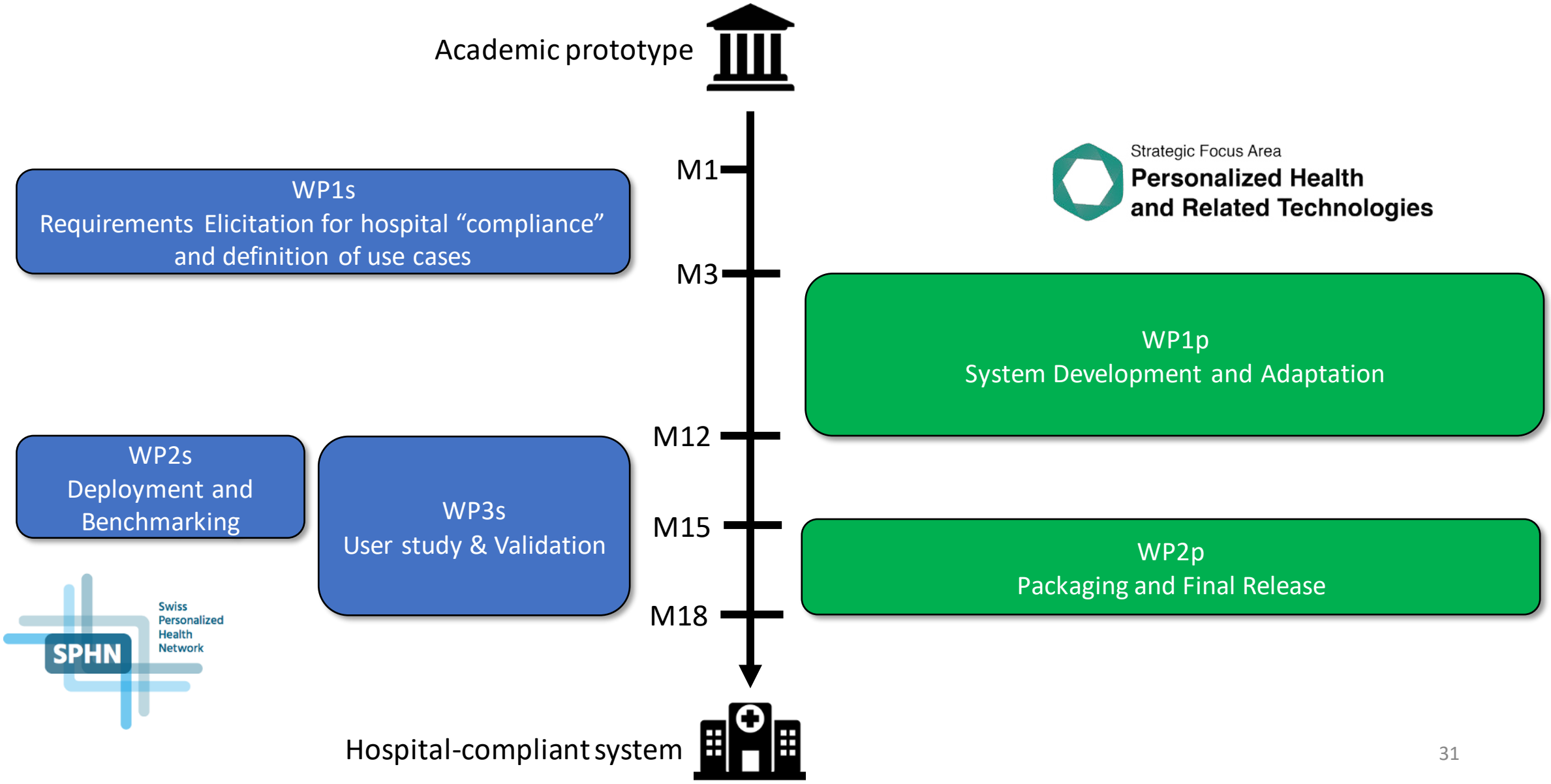
D. Cavin

M. Kämpf

JP Hubaux

LDS, C4DT

Work packages and timeline



MedCo software stack: combining the best of medical informatics and information security

Data model

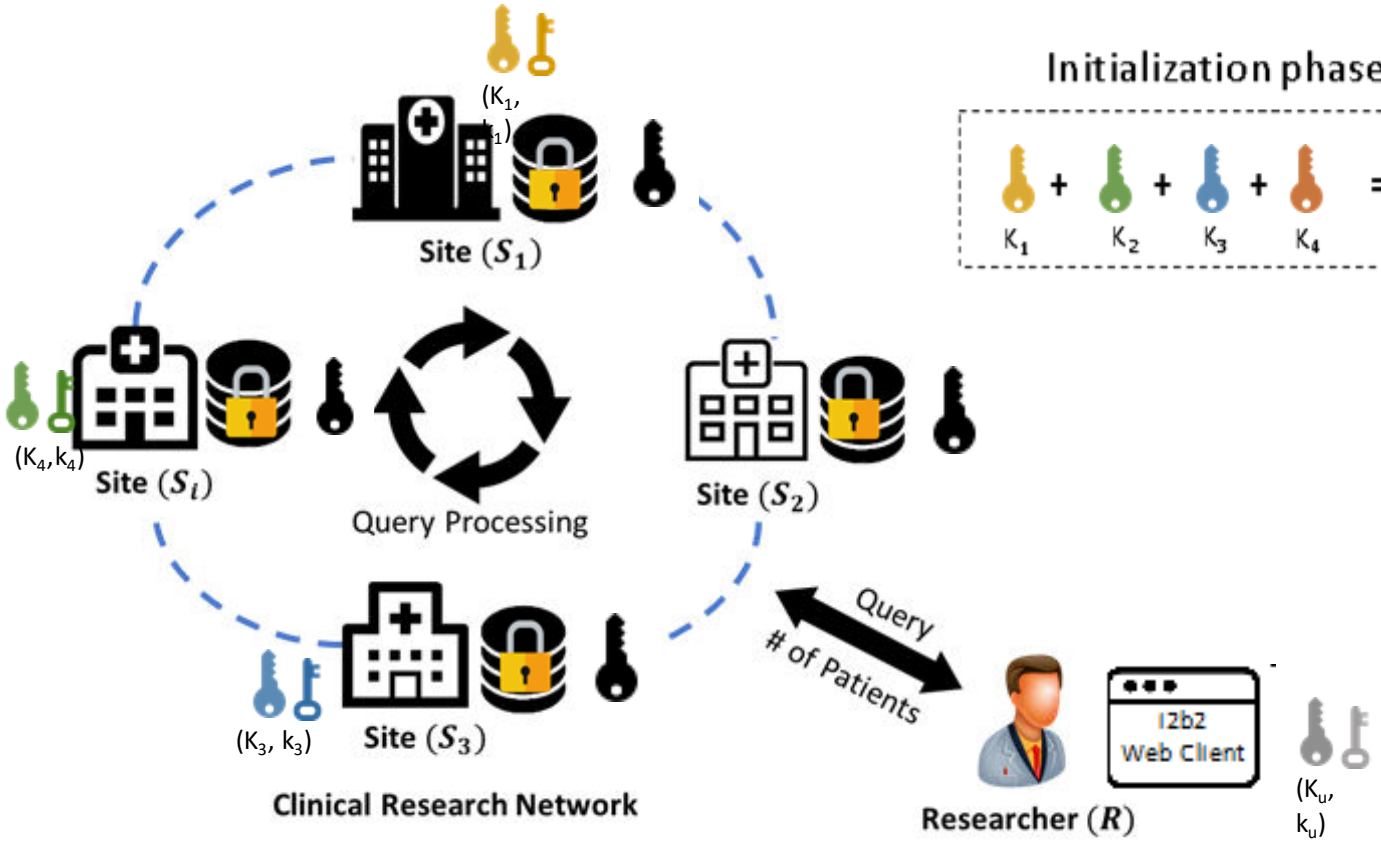
Interoperability layer
Meta API

Privacy-preserving computing framework

Modern GUI

MedCo
Collective protection of medical data

MedCo-Discovery secure query protocol



A, B) ETL & Encryption Phase

1)(user) Query Generation

2)(distributed) Query Tagging

3)(local/dist.) Query Processing

4)(local) Result Aggregation

5)(local) Result Obfuscation

6)(distributed) Results Shuffling

7)(distributed) Results Re-Encryption

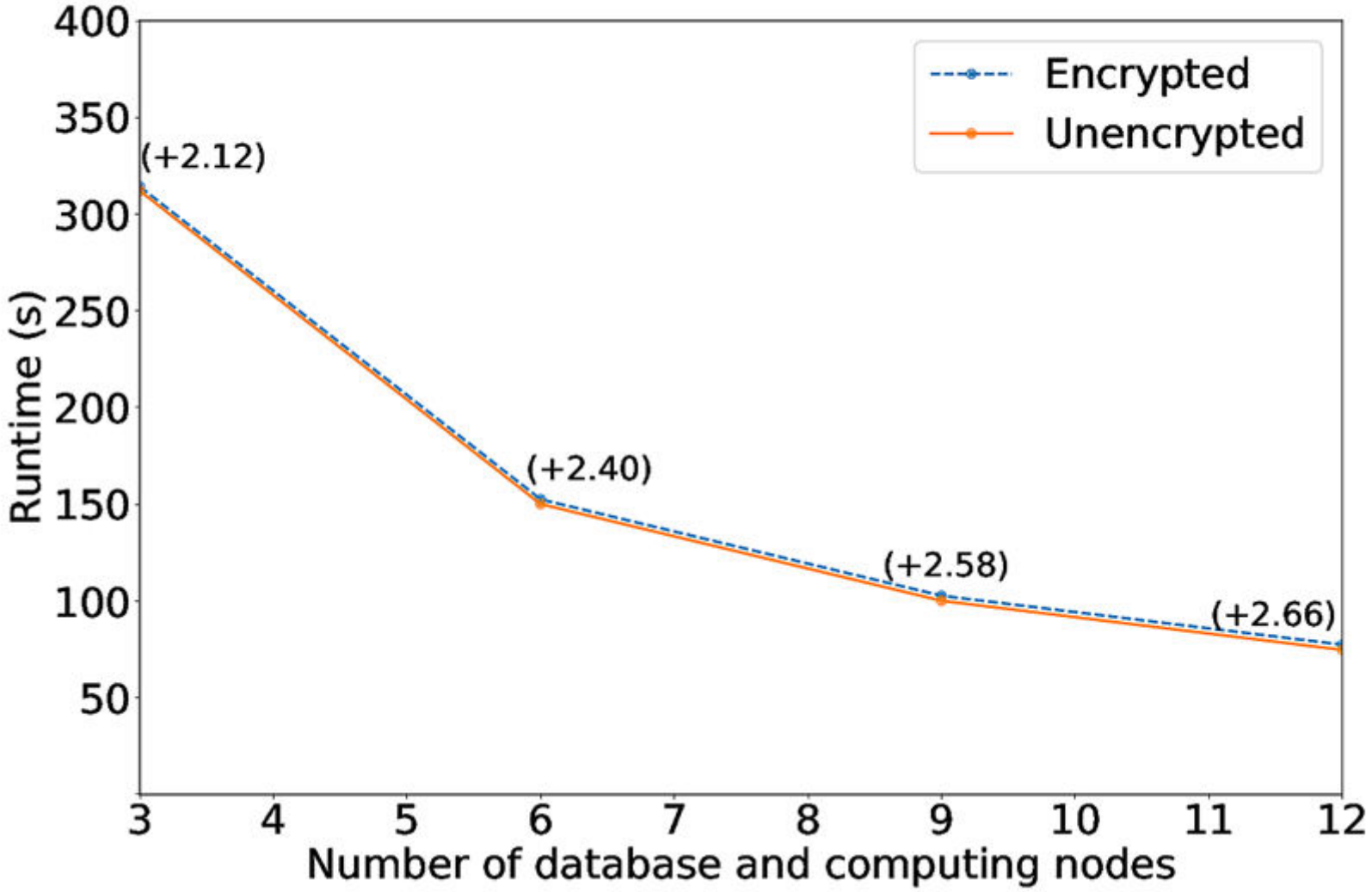
8)(user) Result Decryption

Sites' public keys

Sites' secret keys

Collective public key

MedCo-Explore scalability tests



Population
150’000 individuals
Observations/individual
(15’000, 200’000)
Dataset size
up to 28 billion observations
Query size
(1,50) terms
Resulting set
(100,1511) individuals/node
#servers
(3,12)

28 B data points
1511 matching patients
10 query terms

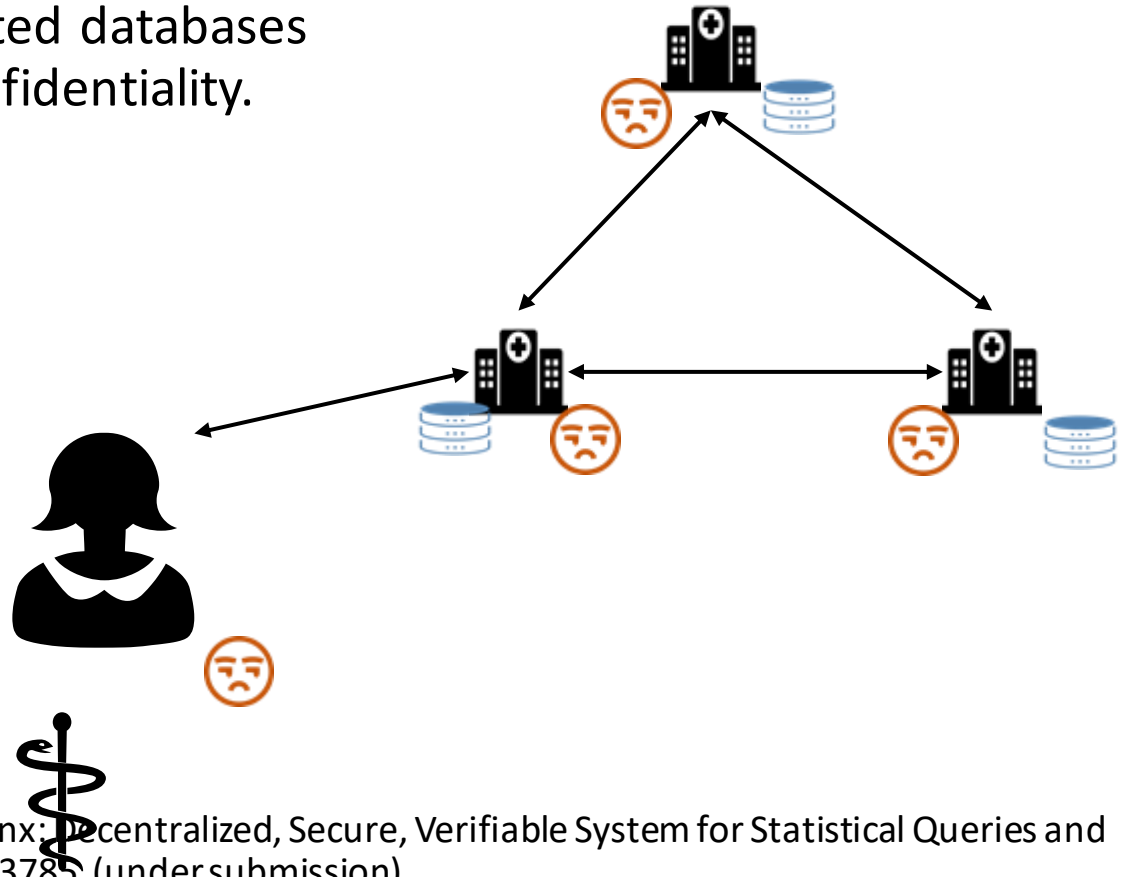
MedCo-Analysis

Decentralized, Secure, Verifiable System for Statistical Queries and Machine Learning on Distributed Databases [1]

Functionality: Enable queries on a set of distributed databases while protecting individuals privacy and data confidentiality.

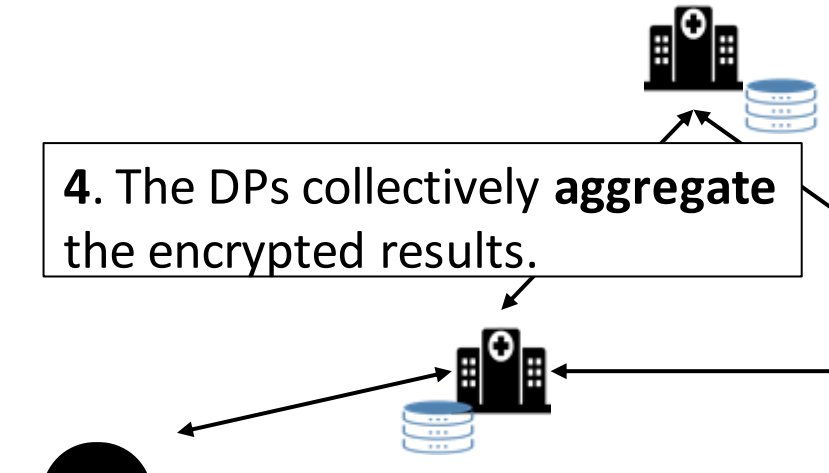
Statistics
sum/count/frequency count
and/or, max/min
variance/standard deviation
Set intersection/union
Cosine similarity

Machine Learning
linear regression
logistic regression



[1] D. Froelicher, J. R. Troncoso-Pastoriza, J. S. Sousa & J.-P. Hubaux. Drynx: Decentralized, Secure, Verifiable System for Statistical Queries and Machine Learning on Distributed Datasets. arXiv preprint arXiv:1902.03785. (under submission)

MedCo-Analysis: Query Workflow



1. The querier defines the **query**: training of a linear regression model on specific attributes

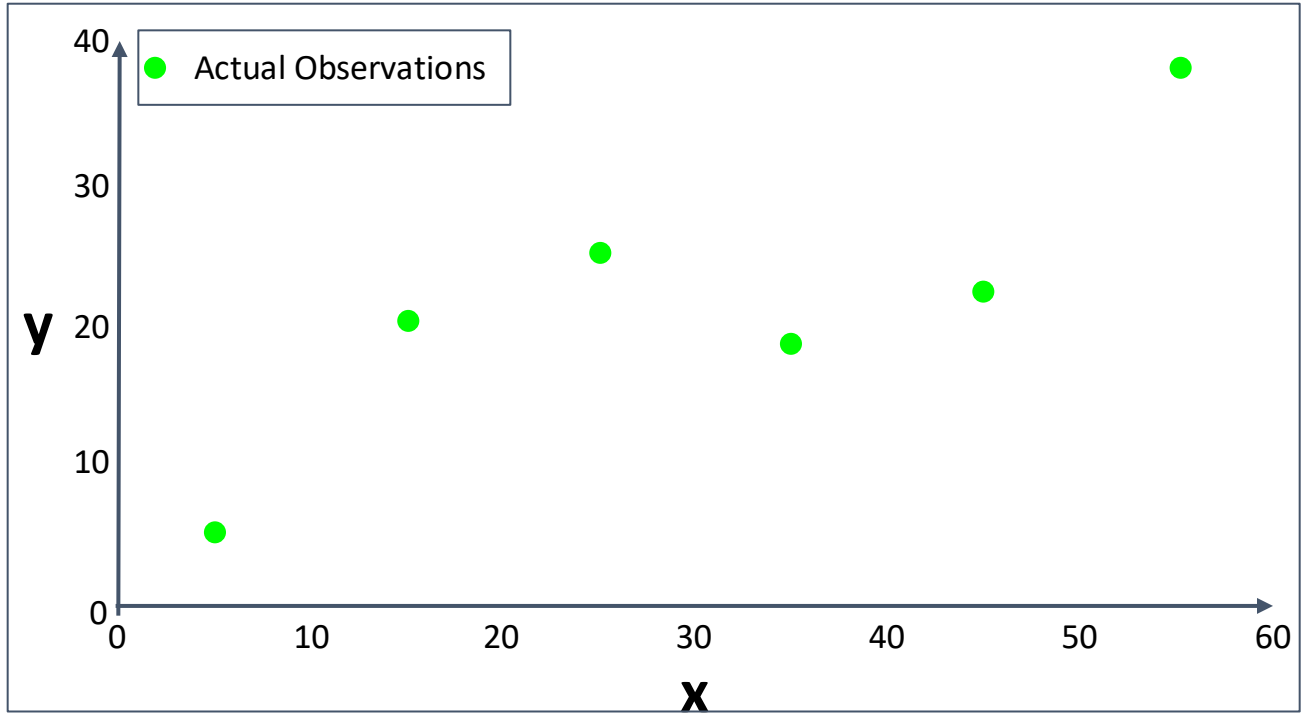
2. Each data provider *i* (DP_i) **locally computes** a function σ on its local database d_i.
3. Each data provider **encrypts** its result.

4. The DPs collectively **aggregate** the encrypted results.

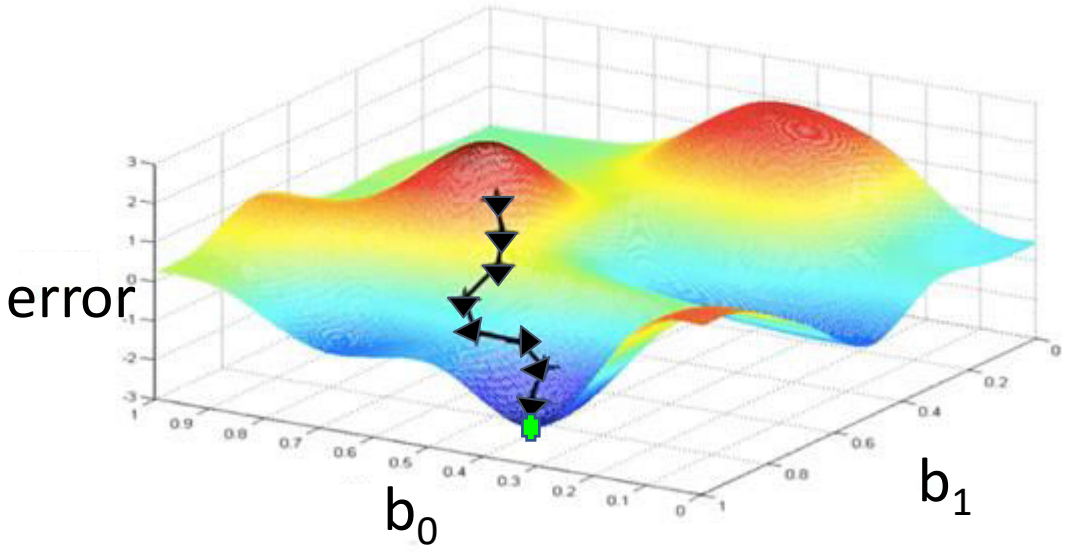
5. The querier can **decrypt and compute** the final result

Linear Regression

Goal: Find the line (defined by $\mathbf{b_0}$ and $\mathbf{b_1}$) that best fits the dots (x_i, y_i).

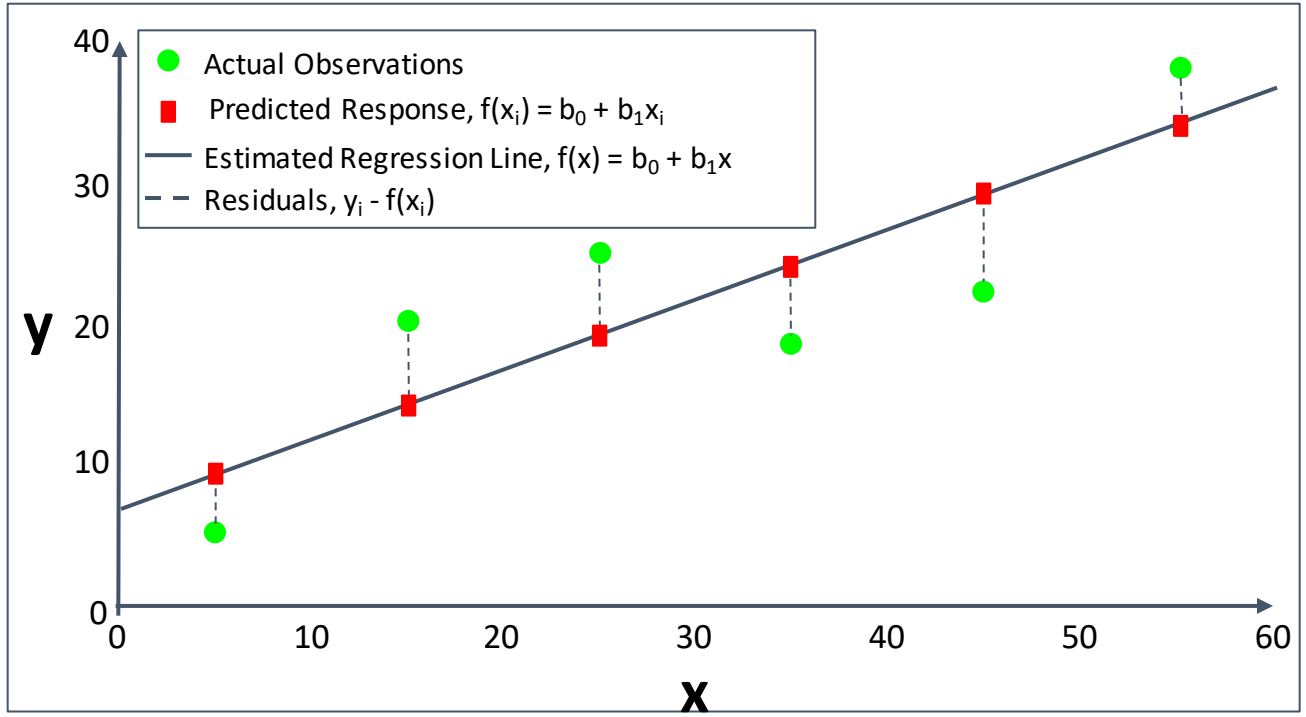


Generic Method to find the best $\mathbf{b_0}$ and $\mathbf{b_1}$: gradient descent is used to find the $\mathbf{b_0}$, $\mathbf{b_1}$ that give the minimum error.

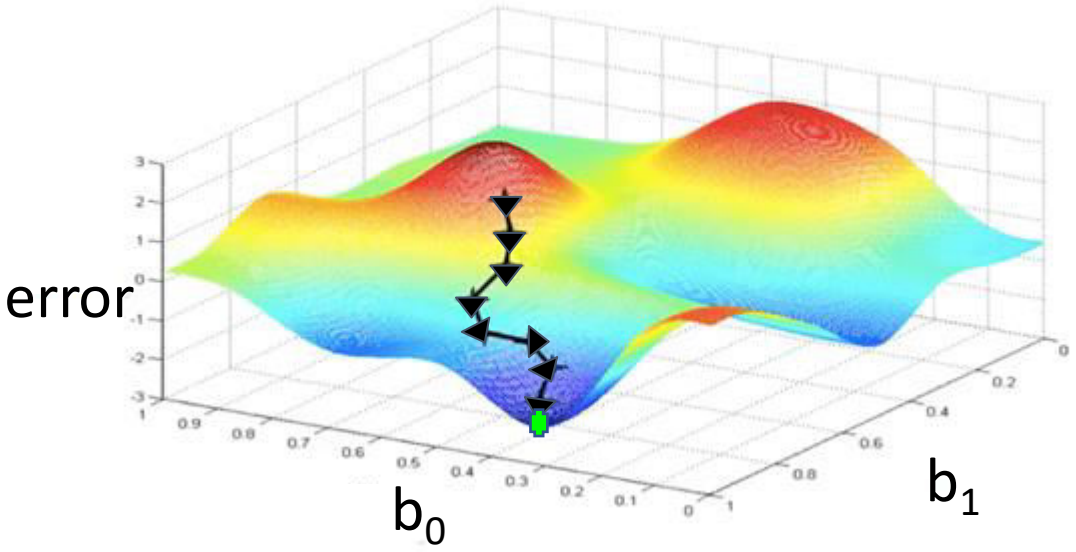


Linear Regression

Goal: Find the line (defined by $\mathbf{b_0}$ and $\mathbf{b_1}$) that best fits the dots (x_i, y_i).

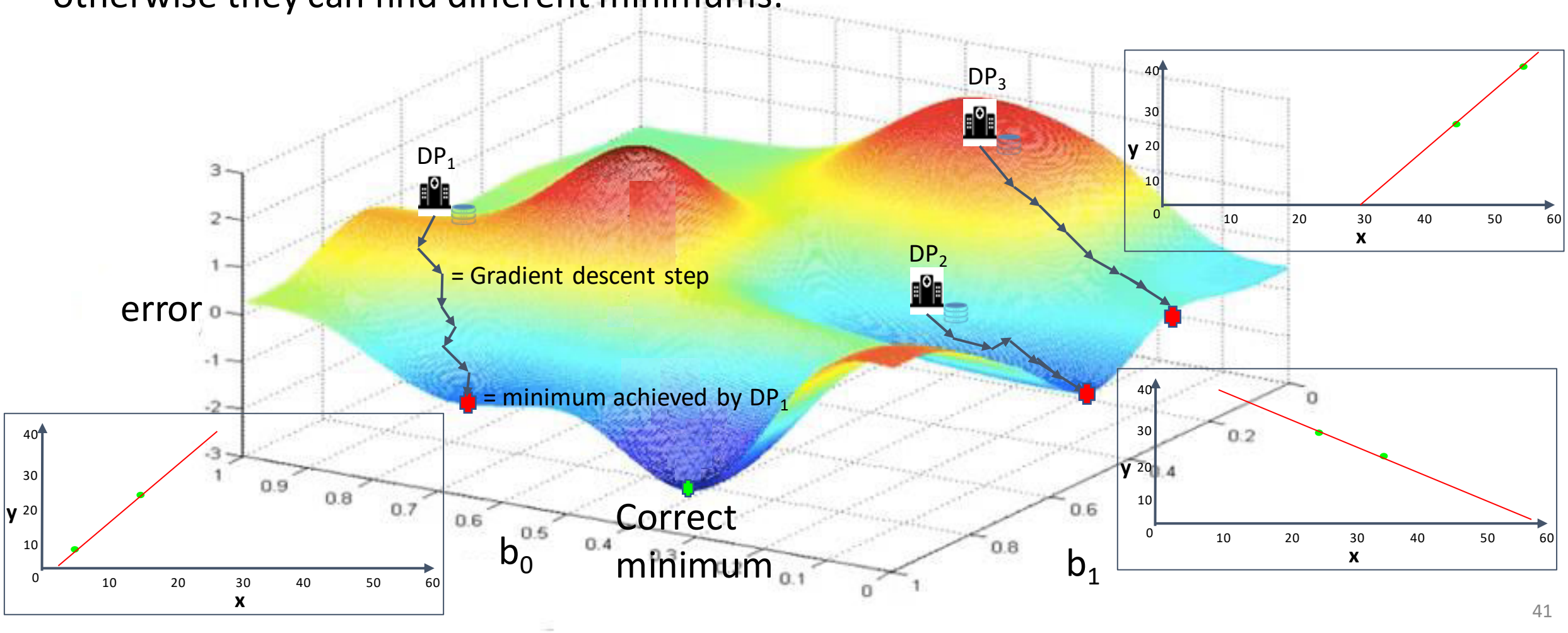


Generic Method to find the best $\mathbf{b_0}$ and $\mathbf{b_1}$: gradient descent is used to find the $\mathbf{b_0}$, $\mathbf{b_1}$ that give the minimum error.



Distributed Linear Regression

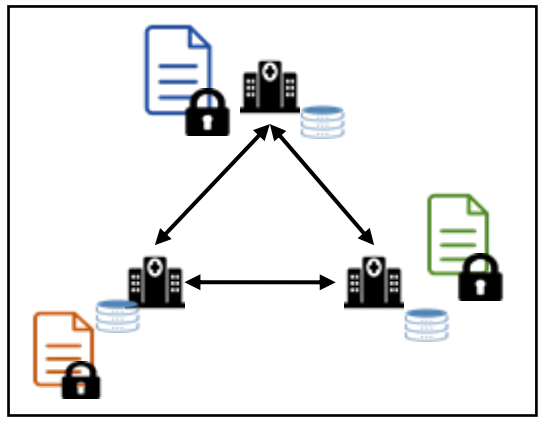
Problem: the data providers have to collaborate during the gradient descent, otherwise they can find different minimums.



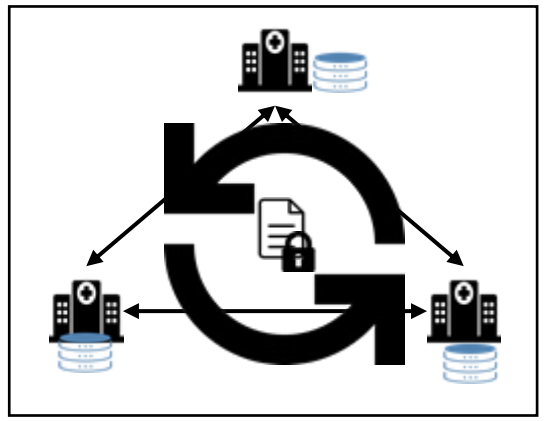
Example: Distributed Linear Regression

Solution: the data providers collaborate to enable a joint gradient descent while protecting their privacy

1. DPs create encrypted summary of their data

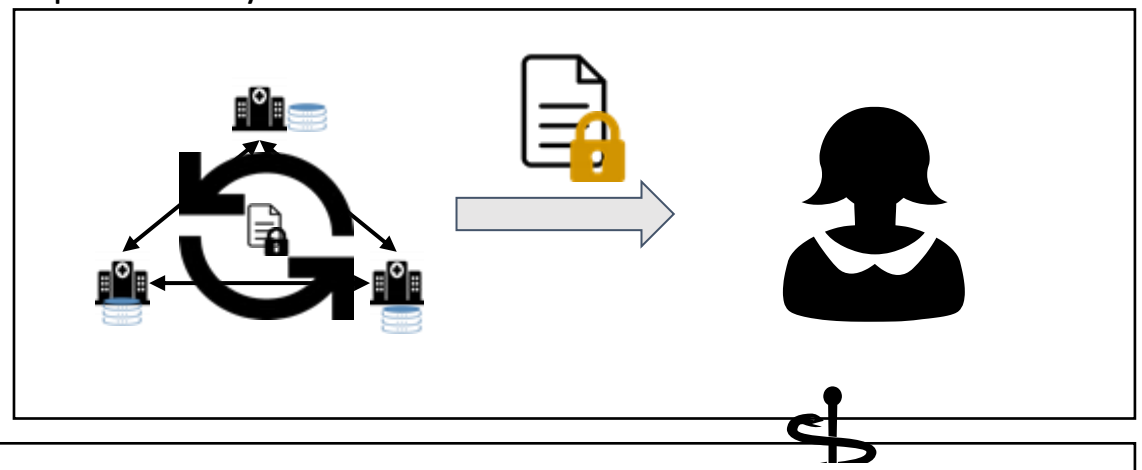


2. DPs' summaries are collectively aggregated

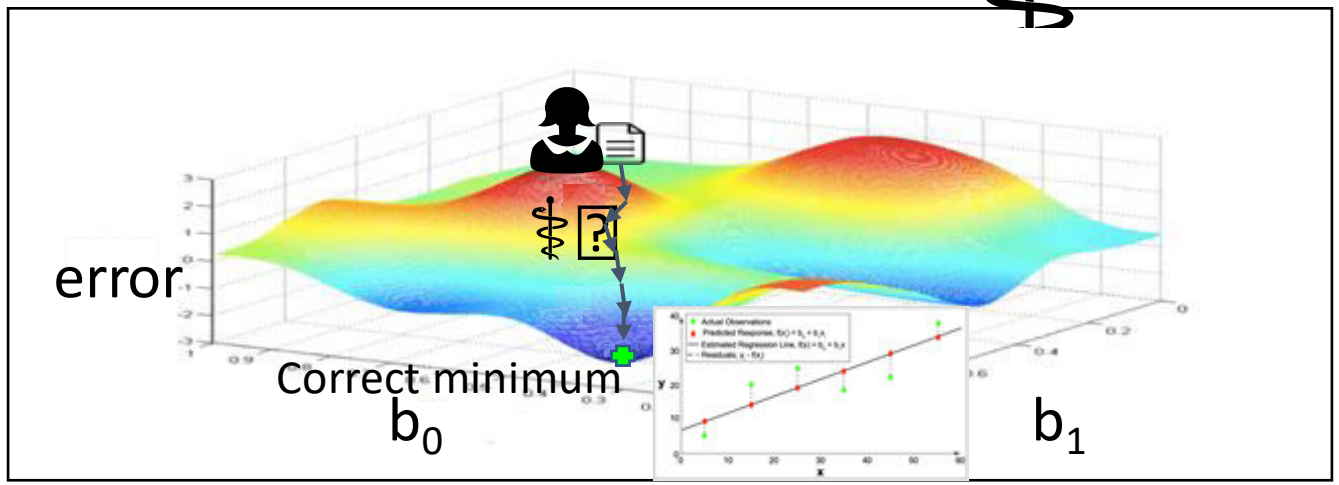


3. The aggregated summary encryption is switched to the querier's key

4. The querier decrypts the final summary

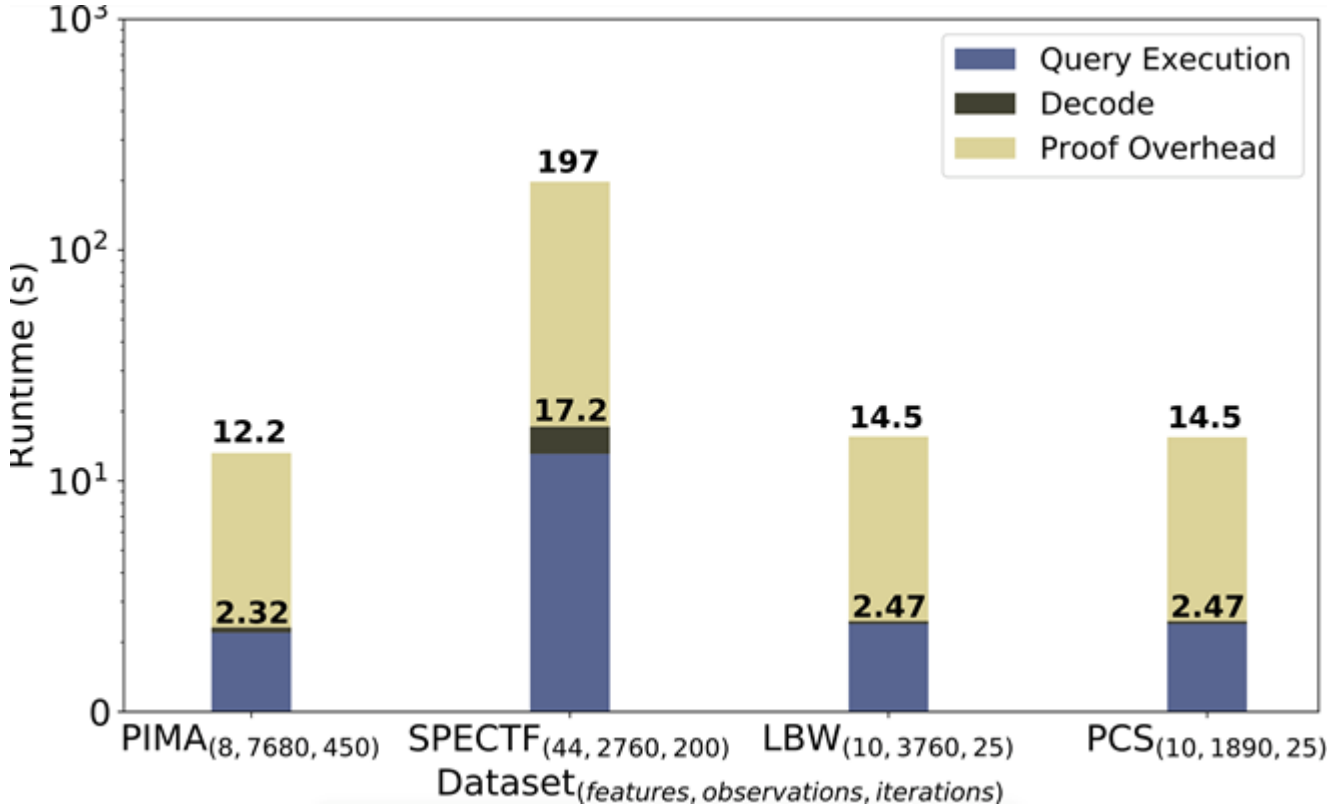


5. The querier performs the gradient descent on the final data summary



Possible technique:
alternating direction method of multipliers (ADMM)
Boyd et al., 2011

Example: Distributed Logistic Regression - Evaluation



Data.		Accuracy
LBW [6]	Centralized	69.31%
	MedCo-Analysis	70.26%
PCS [12]	Centralized	74.60%
	MedCo-Analysis	75.13%
Pima [10]	Centralized	80.5%
	MedCo-Analysis	77.55%
SPECTF [13]	Centralized	78.9%
	MedCo-Analysis	74.87%

LBW = Low birth weight dataset. 10 feat. http://course1.winona.edu/bdeppa/Biostatistics/Data%20Sets/lowbirtharc.txt

PCS = Prostate Cancer Study. 10 feat. http://course1.winona.edu/bdeppa/Biostatistics/Data%20Sets/Prostate%20Logistic.txt

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

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

Parameters:
6 Computing Nodes, 7 Verifying Nodes
60 DPs
80% training; 20% testing
Scaling factor 10^2;
learning rate 0.1;
k = 2;
l2-regularization factor = 1;

MedCo Features and Guarantees

Functionalities

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

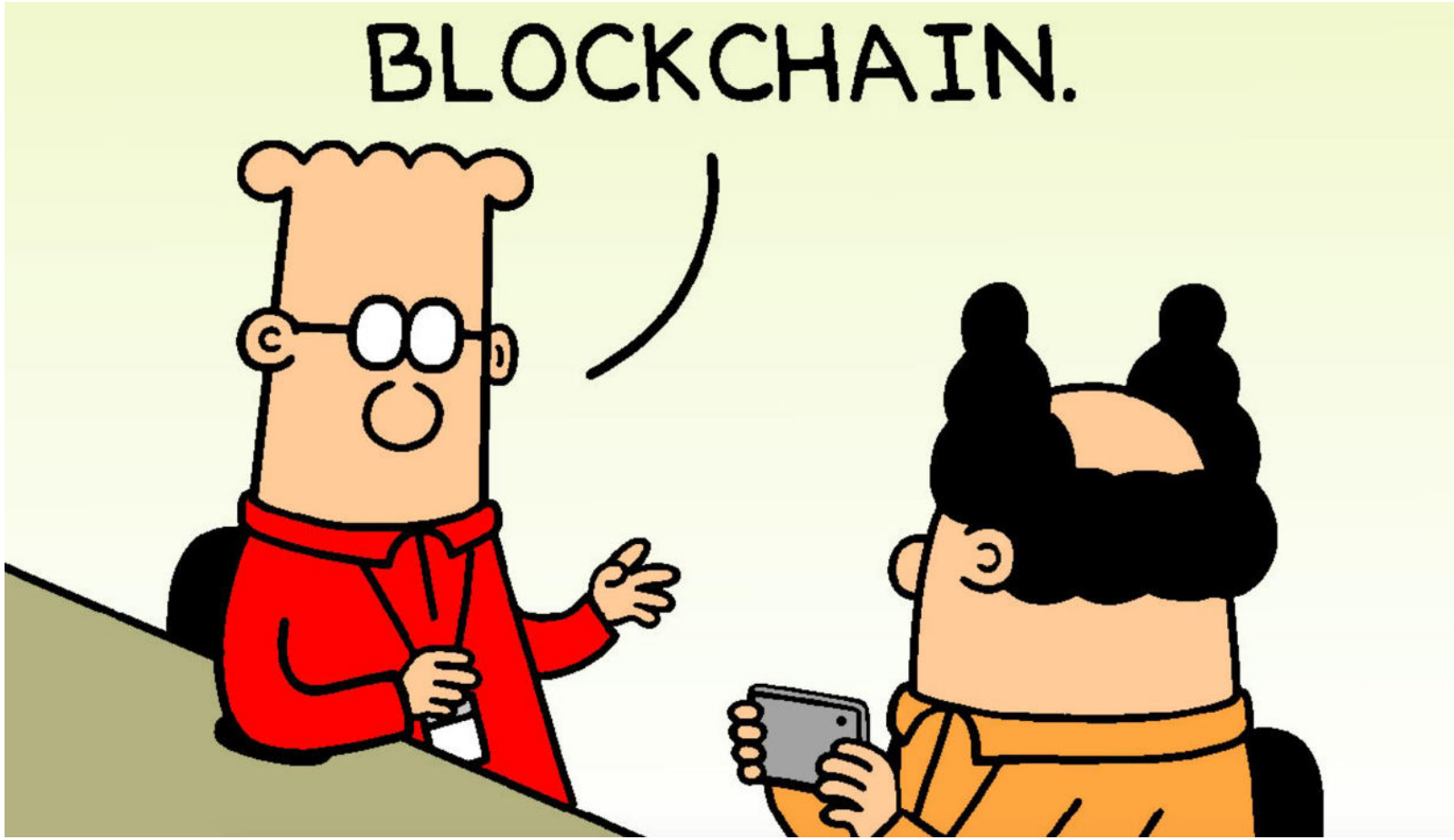
linear regression
logistic regression
...

Data Confidentiality

The data never leave the data providers' premises.

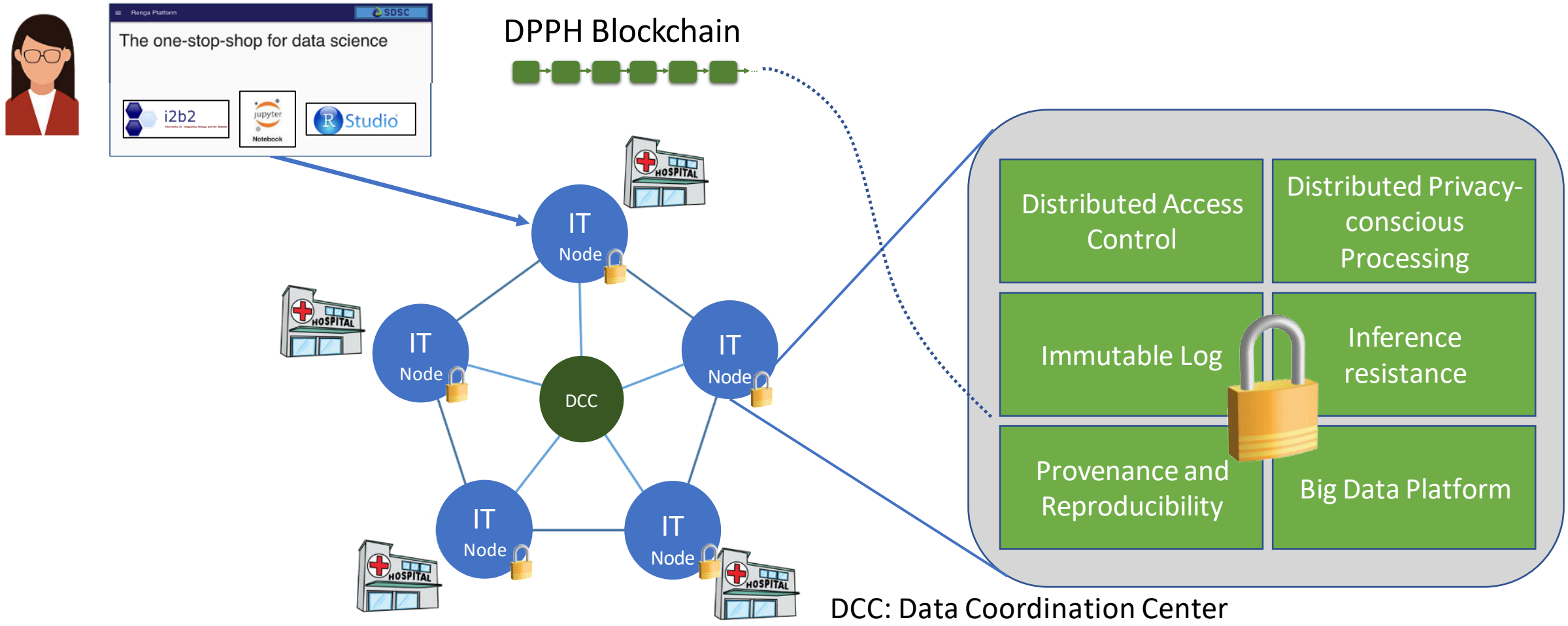
Privacy

The querier only sees the final result aggregated among multiple data providers.



© Dilbert.com

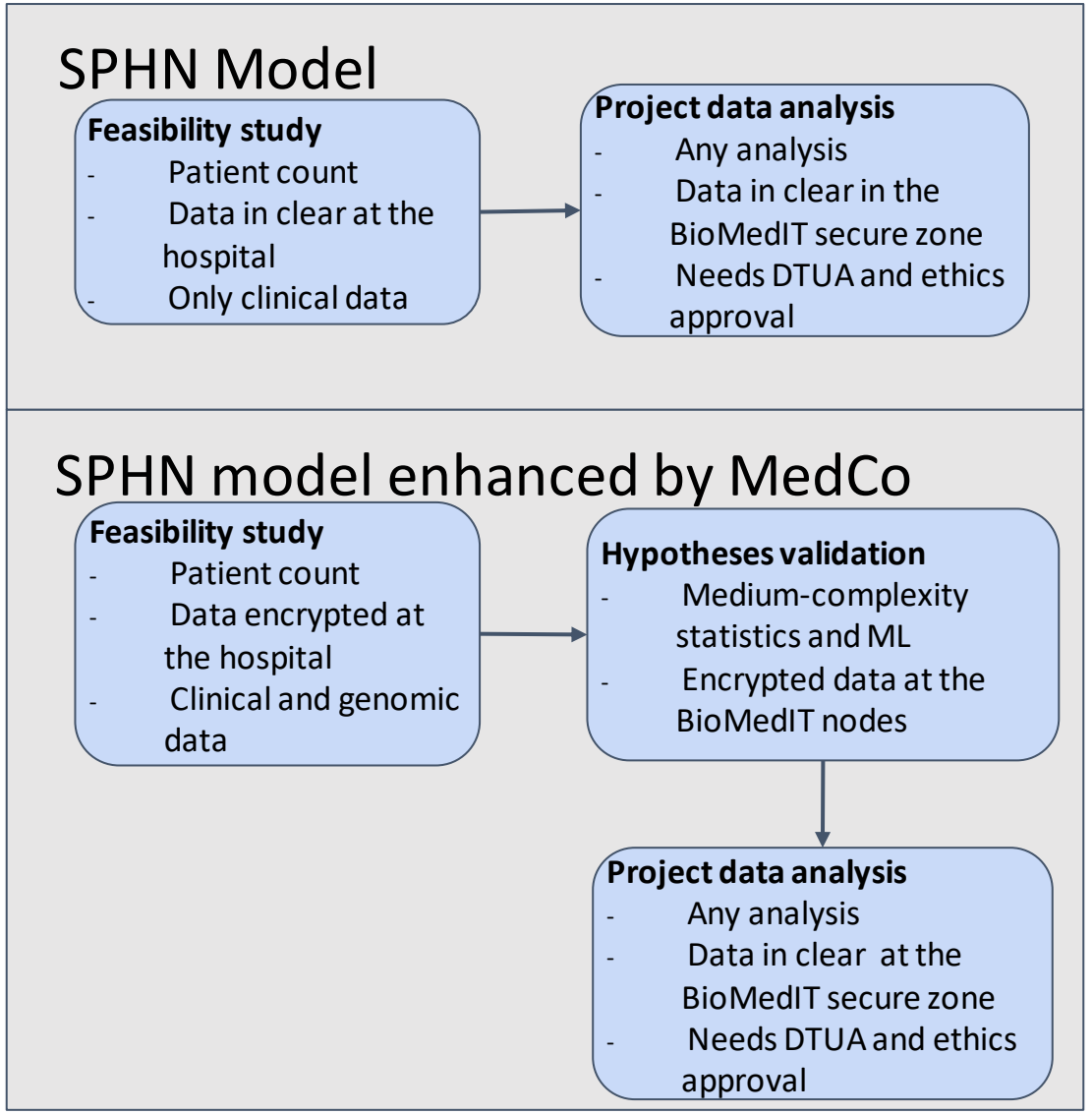
DPPH – The Role of the Blockchain



We use a **closed** ("permissioned") blockchain, unlike Bitcoin that uses a **public** ("non-permissioned") blockchain.

MedCo: Alleviating Data Access for Researchers

- **End-to-end and collective protection** of patient individual-level data ⇒ nobody has access to data in the clear and researchers only obtain aggregate statistics
- Researchers can perform **low- to medium-complexity analyses** (e.g., correlation analysis, survival analysis, linear/logistic regressions) to **validate their research hypotheses <u>BEFORE</u> launching the administrative process to access data in the clear**
- Similarly to the "feasibility" study phase, **access to the system could be granted to the whole SPHN community** on a **tiered-based** access mode ⇒ this would significantly accelerate research as researchers could quickly refine their study criteria **<u>BEFORE</u>** requesting the access to the data



Post-Quantum Resistance: The Lattigo Library

Lattigo unleashes the potential of **lattice-based cryptography** in **secure multiparty computation** for modern software stacks

Pure Go solution:

- Modern language
- Fast & Memory safe
- Ease of build

Lattice-based cryptography:

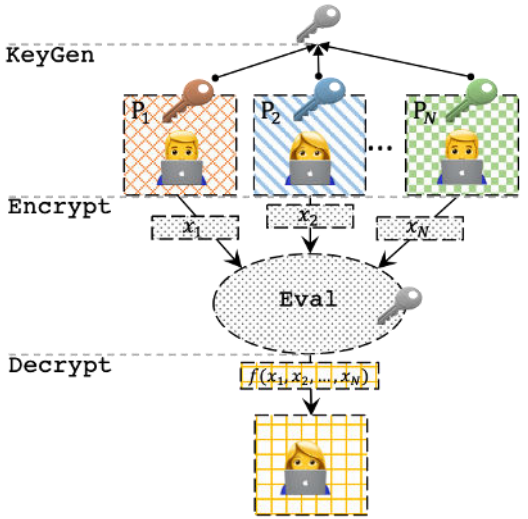
- Post-quantum security
- Fast algorithms
- Versatile constructions

Homomorphic encryption:

- Encrypted integer-arithmetic
- Encrypted complex/float-arithmetic
- Distributed cryptosystems

Secure Multiparty Computation:

- Decentralization
- Secure data sharing



Upcoming support for:

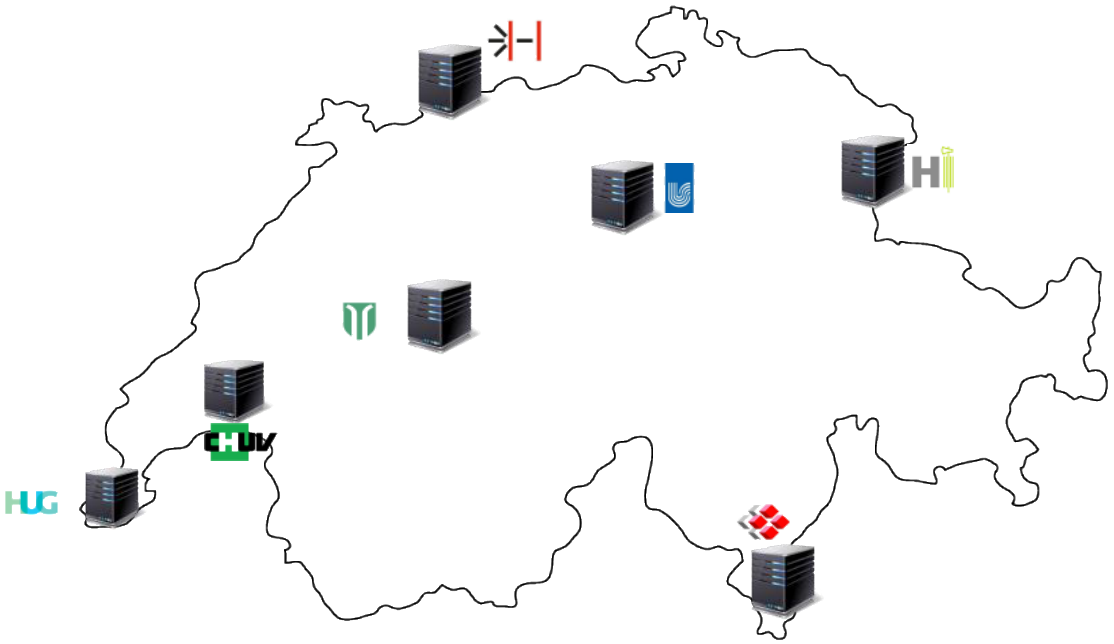
Fully homomorphic encryption

Post-quantum key exchange

General purpose SMC Engine

https://lattigo.epfl.ch

How about the other 99.9% Human Beings?



Swiss Personalized Health Network

At the international level:

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

GA4GH has its own workstream on data security

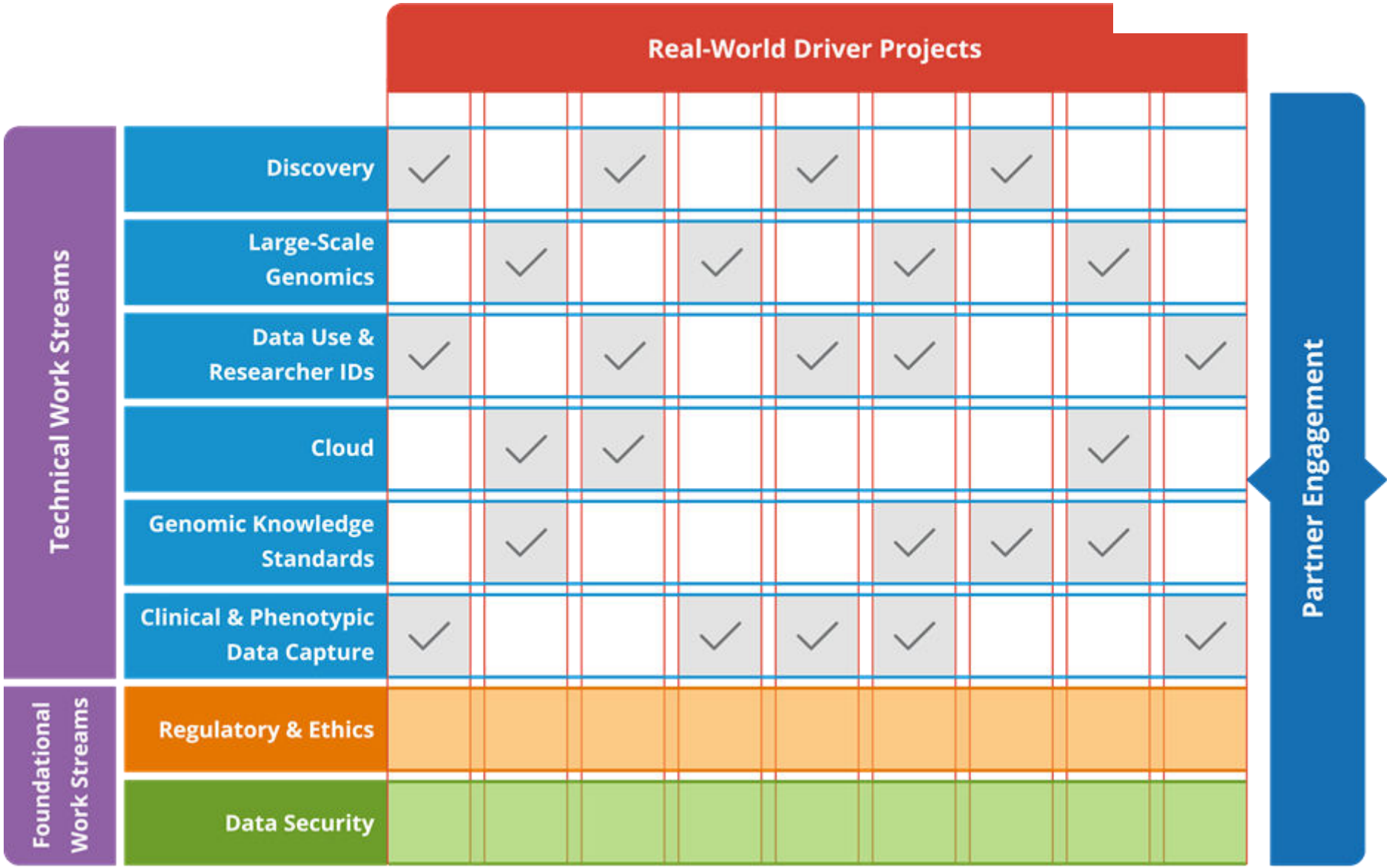
World Wide Web of –omic and Health Data

2013: creation of the **Global Alliance for Genomics & Health**

GA4GH Organization

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



Work Streams vs Driver Projects

Global Alliance for Genomics & Health

Work Streams

- Internal to GA4GH
- Deliver standards and policy frameworks based on the Strategic Roadmap
- Run by 2 volunteer Leads within the community
- Contributors come from a variety of projects and organizations

Example:

Data Use and Researcher Identities

Driver Projects

- External to GA4GH
- Provide input towards the Strategic Roadmap and standards development
- Contribute resources to Work Streams for standards development
- Pilot implementations for new standards

Example:

ARGO Accelerating Research in Genomic Oncology

International Cancer Genome Consortium

Data Security

Global Alliance for Genomics & Health

Technology standards and best practices for protecting data

- **Authentication and authorization infrastructure (AAI)**: GA4GH standard technical profile for authenticating the identity of individuals seeking to access data and services
- **Breach Response Protocol:** protocol for the GA4GH community to effectively respond to and recover from security breaches
- **Ongoing discussions**
  - on homomorphic encryption and SMC

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)
  - **October 21-22, 2019, Boston (co-located with GA4GH)**
- **iDash**: integrating Data for Analysis, Anonymization and sHaring (annual event)
- 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

➔ Lots of material online

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

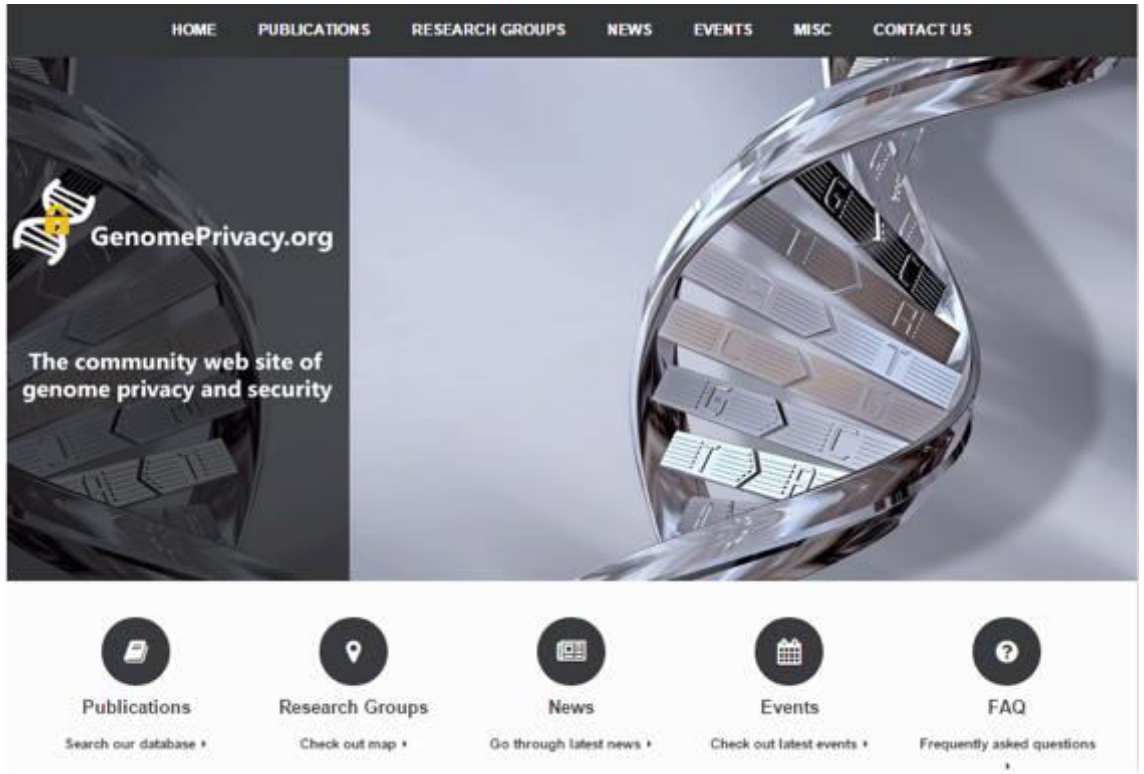
iDASH

ipam

DPPH18

DPPH18.epfl.ch

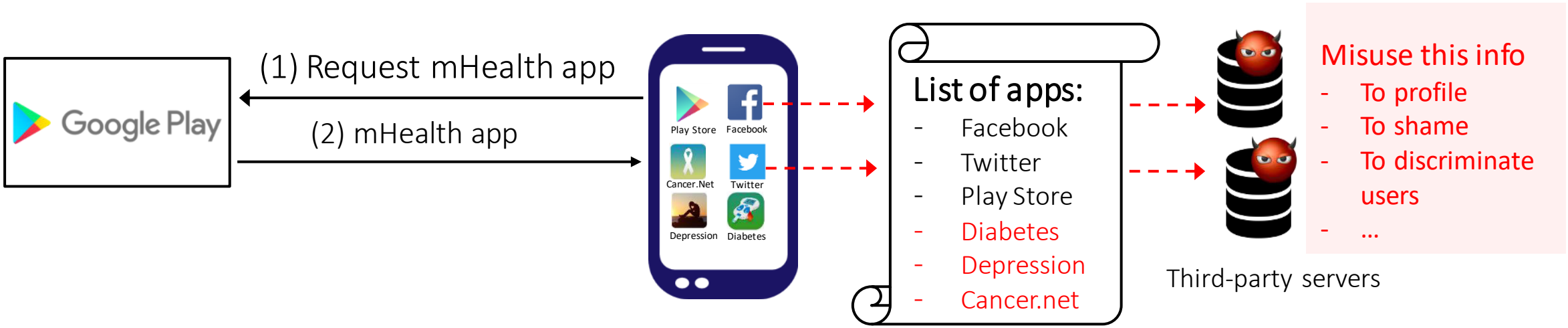
genomeprivacy.org



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)

Privacy Challenge in mHealth

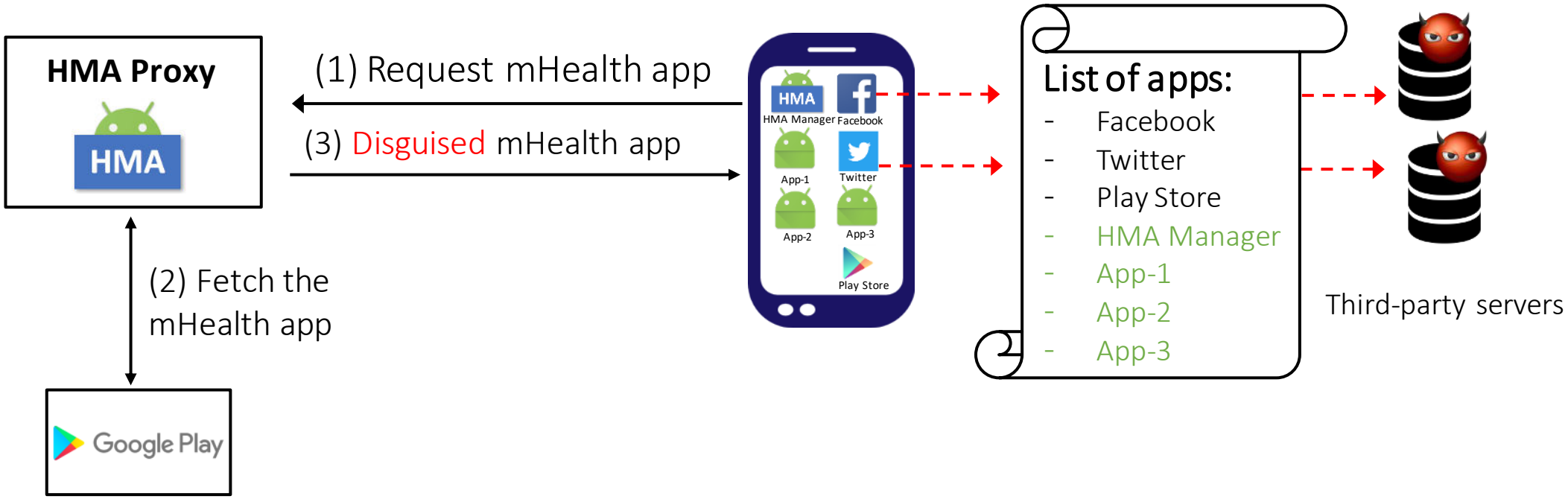


- Many apps collect the list of installed apps
- Presence of an mHealth app → specific medical conditions of its users
- Collected lists of installed apps can be shared with third-parties

How to hide the presence of a sensitive app from other apps while preserving key functionalities and usability of the app, and without requiring users to modify the OS of their phones?

Our solution: HideMyApp (HMA)

- Main idea: Launch the sensitive app <u>without</u> installing it



- Technologies used:
  - Dynamic loading of classes and resources from an application package (APK)
  - App virtualization
  - Randomization and obfuscation

https://hma.epfl.ch

Conclusion

- Protecting health data is one of the most formidable **cybersecurity challenges**
- What is at stake is no less than **human dignity** and **democracy**
- With the advent of **molecular medicine** (including genomics):
  - risk is increasing
  - conventional medical data protection techniques based on de-identification do not work anymore
- **Distributed cohorts** will play a key role
- **Solutions** will be technical (crypto, security, statistics,…), legal and organizational

https://dpph.ch

https://medco.epfl.ch