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)
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
17.09.2019, 07:02 Uhr

Millionenfach Patientendaten ungeschützt im Netz

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

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
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
The Genomic Avalanche Is Coming…
The massive digitalization of clinical and genomic information is providing unprecedented opportunities for improvements in diagnosis, preventive medicine and targeted therapies.
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.
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...**
Direct-to-Consumer Genomics (1/2)

- Ancestry.com (millions of customers)
Direct-to-Consumer Genomics (2/2)

- 23andMe.com (millions customers)

<table>
<thead>
<tr>
<th>Name</th>
<th>Confidence</th>
<th>Your Risk</th>
<th>Avg. Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial Fibrillation</td>
<td>★★★★</td>
<td>33.9%</td>
<td>27.2%</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>★★★★</td>
<td>29.3%</td>
<td>17.8%</td>
</tr>
<tr>
<td>Alzheimer's Disease</td>
<td>★★★★</td>
<td>14.2%</td>
<td>7.2%</td>
</tr>
<tr>
<td>Age-related Macular Degeneration</td>
<td>★★★★</td>
<td>11.1%</td>
<td>6.5%</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>★★★★</td>
<td>7.8%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>★★★★</td>
<td>4.2%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Restless Legs Syndrome</td>
<td>★★★★</td>
<td>2.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Parkinson's Disease</td>
<td>★★★★</td>
<td>2.2%</td>
<td>1.6%</td>
</tr>
</tbody>
</table>
With genetic testing, I gave my parents the gift of divorce

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

<table>
<thead>
<tr>
<th>Traditional Encryption</th>
<th>Homomorphic Encryption</th>
<th>Secure Multiparty Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Protects data at rest and in transit</td>
<td>- Protects computation in untrusted environments</td>
<td>- Protects computation in distributed environments</td>
</tr>
<tr>
<td>- Cannot protect computation</td>
<td>- Limited versatility vs efficiency</td>
<td>- High communication overhead</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trusted Execution Environments</th>
<th>Differential Privacy</th>
<th>Distributed Ledger Technologies (Blockchains)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Protects computation with Hardware Trusted Element</td>
<td>- Protects released data from inferences</td>
<td>- Strong accountability and traceability in distributed environments</td>
</tr>
<tr>
<td>- Requires trust in the manufacturer, vulnerable to side-channels</td>
<td>- Degraded data utility (privacy-utility tradeoff)</td>
<td>- Usually no data privacy</td>
</tr>
</tbody>
</table>
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

Data in cleartext

Secured pipe (that’s 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 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
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

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 \( \mathcal{P} = \{P_1, P_2, \ldots, P_N\} \) would like to compute a function \( f(x_1, x_2, \ldots, x_N) = (y_1, y_2, \ldots, y_N) \) 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

68 MCHF
2017 - 2021

50 MCHF
2017 - 2021

www.sphn.ch

https://www.sphn.ch

https://www.sfa-phrt.ch/
• 4 research groups across the ETH domain + SDSC (Swiss Data Science Center)
• Funding: 3 Millions CHFr$s
• 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
Q1: How many patients with BRCA1 and breast cancer?

Q2: What is the survival rate for cancer patients undergoing a given chemotherapy?
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
Work packages and timeline

Academic prototype

WP1s
Requirements Elicitation for hospital “compliance” and definition of use cases

WP2s
Deployment and Benchmarking

WP3s
User study & Validation

WP1p
System Development and Adaptation

WP2p
Packaging and Final Release

Hospital-compliant system
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-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

Initialization phase

\[ K = K_1 + K_2 + K_3 + K_4 \]

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

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 \( \sigma \) 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 $b_0$ and $b_1$) that best fits the dots $(x_i, y_i)$.

**Generic Method to find the best $b_0$ and $b_1$:** gradient descent is used to find the $b_0, b_1$ that give the minimum error.
**Linear Regression**

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

**Generic Method to find the best $b_0$ and $b_1$:**

Gradient descent is used to find the $b_0$, $b_1$ 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

PCS = Prostate Cancer Study. 10 feat. [12] http://course1.winona.edu/bdeppa/Biostatistics/Data%20Sets/Prostate%20Logistic.txt

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

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.
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 **BEFORE** 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 **BEFORE** requesting the access to the data

**SPHN Model**

- **Feasibility study**
  - Patient count
  - Data in clear at the hospital
  - Only clinical data

- **Project data analysis**
  - Any analysis
  - Data in clear in the BioMedIT secure zone
  - Needs DTUA and ethics approval

**SPHN model enhanced by MedCo**

- **Feasibility study**
  - Patient count
  - Data encrypted at the hospital
  - Clinical and genomic data

- **Hypotheses validation**
  - Medium-complexity statistics and ML
  - Encrypted data at the BioMedIT nodes

- **Project data analysis**
  - Any analysis
  - Data in clear at the BioMedIT secure zone
  - Needs DTUA and ethics approval
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?

At the international level:

Swiss Personalized Health Network

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

## Real-World Driver Projects

<table>
<thead>
<tr>
<th>Technical Work Streams</th>
<th>Discovery</th>
<th>Large-Scale Genomics</th>
<th>Data Use &amp; Researcher IDs</th>
<th>Cloud</th>
<th>Genomic Knowledge Standards</th>
<th>Clinical &amp; Phenotypic Data Capture</th>
<th>Regulatory &amp; Ethics</th>
<th>Data Security</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partner Engagement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Work Streams vs Driver Projects

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

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:
Data Use and Researcher Identities
Data Security

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
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 **without** installing it

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

---

[Diagram showing the process of launching an app using HideMyApp (HMA)]

- **List of apps:**
  - Facebook
  - Twitter
  - Play Store
  - HMA Manager
  - App-1
  - App-2
  - App-3

- **Third-party servers**

---

[Website link: 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://dpdh.ch
https://medco.epfl.ch