ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

IPAM

10-12 January 2018

Security and Privacy of P4 Medicine: Challenges and Possible Solutions

Jean-Pierre Hubaux

With gratitude to the many biomed and CS researchers
with whom I have been fortunate to collaborate on this topic

Growing Concern: Medical Data Breaches

Around 1 declared breach per day, 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
	East Central Kansas Area Agency on Aging	KS	Business Associate	8750	10/31/2017	Hacking/IT Incident	Network Server
	Texas Children's Health Plan	TX	Health Plan	932	10/27/2017	Unauthorized Access/Disclosure	Email
	Catholic Charities of the Diocese of Albany	NY	Healthcare Provider	4624	10/27/2017	Hacking/IT Incident	Network Server
	Arch City Dental, LLC - Drs. Baloy and Donatelli	OH	Healthcare Provider	1716	10/26/2017	Unauthorized Access/Disclosure	Email
	MGA Home Healthcare Colorado, Inc.	AZ	Healthcare Provider	2898	10/25/2017	Hacking/IT Incident	Email
	TJ Samson Community Hospital	KY	Healthcare Provider	683	10/24/2017	Unauthorized Access/Disclosure	Electronic Medical Record
	Brevard Physician Associates	FL	Healthcare Provider	7976	10/24/2017	Theft	Desktop Computer
	Aetna, Inc.	CT	Health Plan	1506	10/23/2017	Unauthorized Access/Disclosure	Network Server
	Recovery Institute of the South East P.A.	FL	Healthcare Provider	689	10/21/2017	Hacking/IT Incident	Desktop Computer, Electronic Medical Record, Email, Laptop, Network Server, Other, Other Portable Electronic Device, Paper/Films

“WannaCry” Ransomware Virus (May 2017)

Do state institutions have the resources to fight hackers?

Public sector has lessons to learn as hospital trusts and GPs struggle to recover from ransomware attack

A ransomware attack bought computers to a standstill across the world on Friday. Photograph: Ritchie B. Tongo/EPA

The Guardian,
14 May 2017

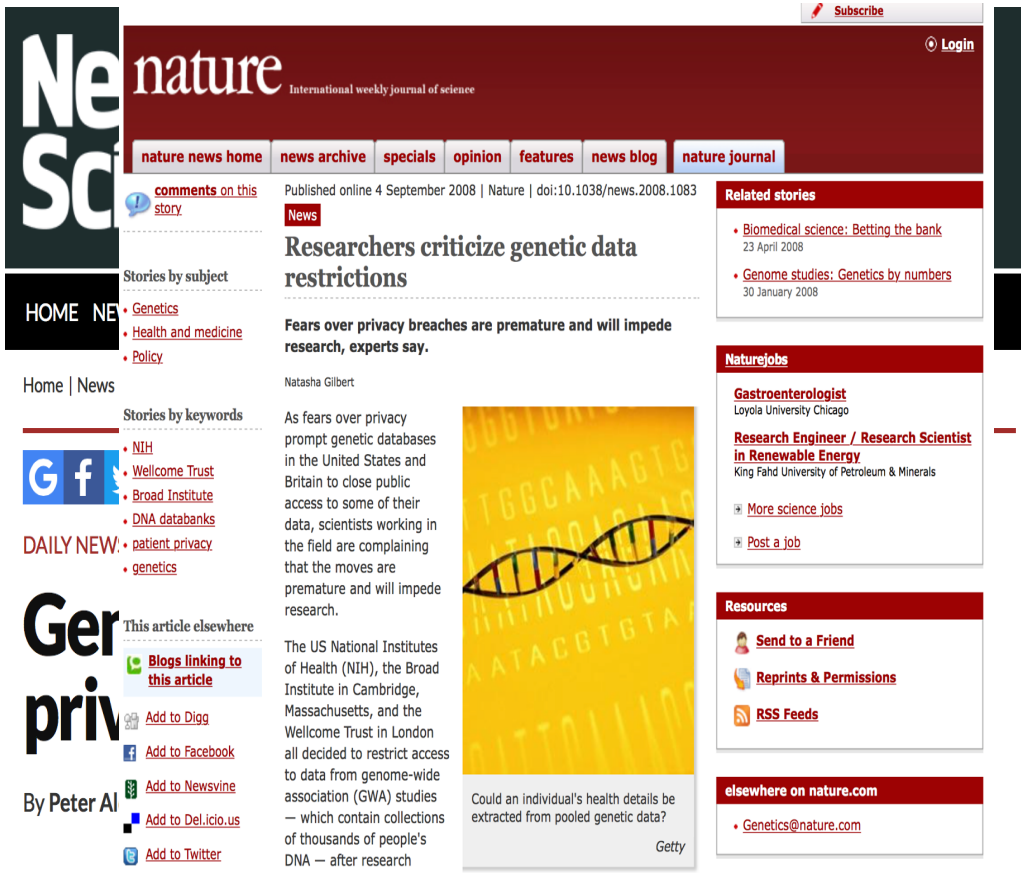
Ransomware Attack against German Hospitals

20. März 2016, 10:05 Uhr Klinikum Neuss

Wenn Cyberkriminelle ein Krankenhaus lahmlegen

Das Lukaskrankenhaus der Städtischen Kliniken in Neuss wurde Opfer von Cyberkriminellen. (Foto: dpa)

Another Major Concern: Re-identification Attacks against Genomic Databases



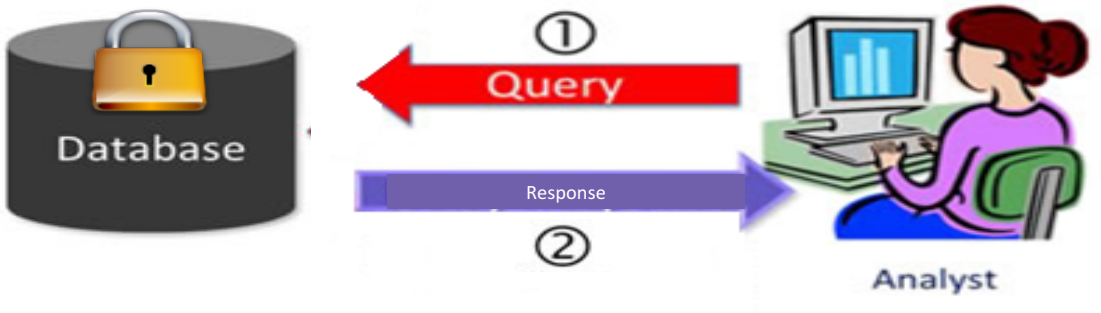
Security / Privacy Requirements for Personalized Health

- Pragmatic approach, **gradual** introduction of new protection tools
- Different **sensitivity levels** of the data
- Different **access rights**
- Exploit **existing** data (electronic health records) and tools
- Be **future-proof** (no short-sighted “bricolage”)
- Awareness of **patient consent**
- Secure also the **collection** of health data (via smartphones, wearable sensors,…)

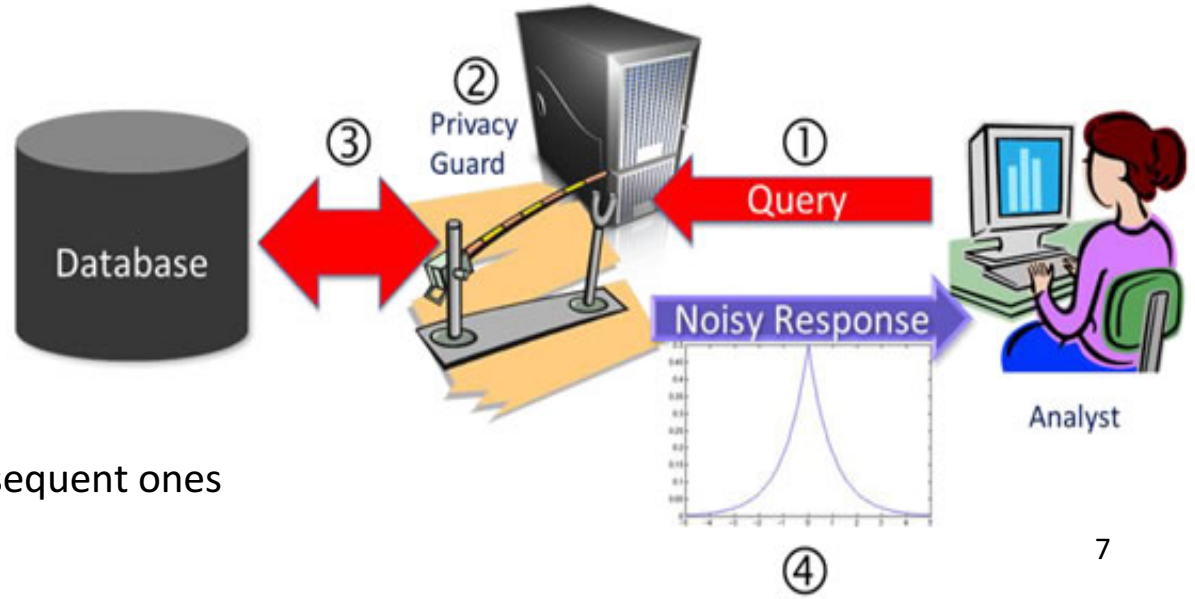
Privacy-Enhancing Technologies

Two main approaches:

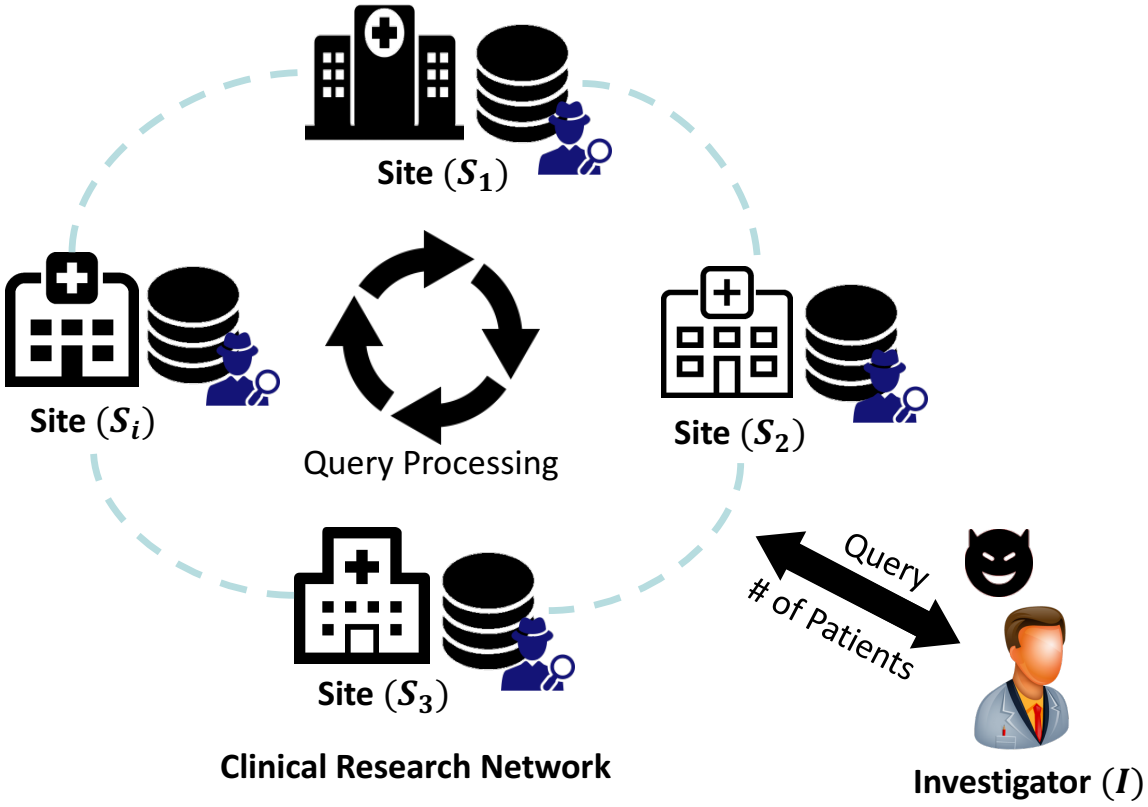
- Protect the data themselves:
  Use of **cryptography**
  - Symmetric / asymmetric encryption
  - Property-preserving encryption
  - (Partially) homomorphic encryption
  - …



- Avoid that responses leak "too much" information:
  Provide only **global**
- (e.g., statistical) **results**
  - K-anonymity, l-diversity, t-closeness
  - Differential privacy
  - For genomics, see "Homer attack" and subsequent ones



MedCo: System and Threat Models



Honest-but-curious adversary:

- honestly follows the protocol
- tries to infer sensitive data from the different steps of the protocol

Malicious-but-covert adversary:

- can tamper with the protocol
- tries to infer sensitive data from the query end-result

Main Concerns

- **Attribute disclosure** due to illegitimate access to the data
  - External (hacker) or internal (insider) attacker stealing the data

➔ Standard encryption can protect data ONLY at rest or in transit BUT NOT during processing (e.g., in the memory)

- **Patient re-identification** due to legitimate access to the data
  - Malicious users performing “smart” data requests in order to re-identify patients in a specific dataset (e.g., patients with HIV)

➔ De-identification or anonymization is ineffective with genomic data

Main Requirements

Functionality:

COUNT(patients**)/SELECT(**patients**)**

FROM database

WHERE * AND/OR *

GROUP BY *

* represents any possible concepts in the ontology

Security/Privacy:

- Protection of data confidentiality at rest, in transit and **during computation**
- no single point of failure
- only the investigator can obtain the query end-result
- (optional) unlinkability
- (optional) differential privacy

MedCo: Combining the Best of Both Worlds

Biomedical Informatics:

- Data model from *i2b2* *(Informatics for Integrating Biology and the Bedside)*

  Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)*. Journal of the American Medical Informatics Association. 2010 Mar 1;17(2):124-30.

- Interoperability layer from *SHRINE*

  McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, Bickel J, Wattanasin N, Gilbert C, Trevvett P, Churchill S. *SHRINE: enabling nationally scalable multi-site disease studies*. PloS one. 2013 Mar 7;8(3):e55811.

IT Privacy and Security:

- Privacy-preserving distributed protocols from *UnLynx*

  Froelicher, D., Egger, P., Sousa, J.S., Raisaro, J.L., Huang, Z., Mouchet, C., Ford, B. and Hubaux, J.P., 2017. UnLynx: A Decentralized System for Privacy-Conscious Data Sharing. In *Proceedings on Privacy Enhancing Technologies* (Vol. 4, pp. 152-170).

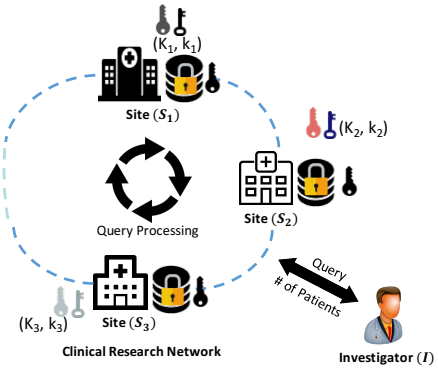
Use Case: Tests on Clinical Oncology

Public Data from cBioPortal

- 121 patients (later scaled to 121,000) with 9 clinical attributes and 1,978 mutations on average per site and patient
- **Query 1**: "*Number of patients with skin cutaneous melanoma AND a mutation in BRAF gene affecting the protein at position 600.*"
  ➔ (2 clinical attributes, 4 mutations)
- **Query 2:** "*Number of patients with skin cutaneous melanoma AND a mutation in BRAF gene AND a mutation in (PTEN OR CDKN2A OR MAP2K1 OR MAP2K2 genes)*"
  ➔ (2 clinical attributes, 77 mutations)

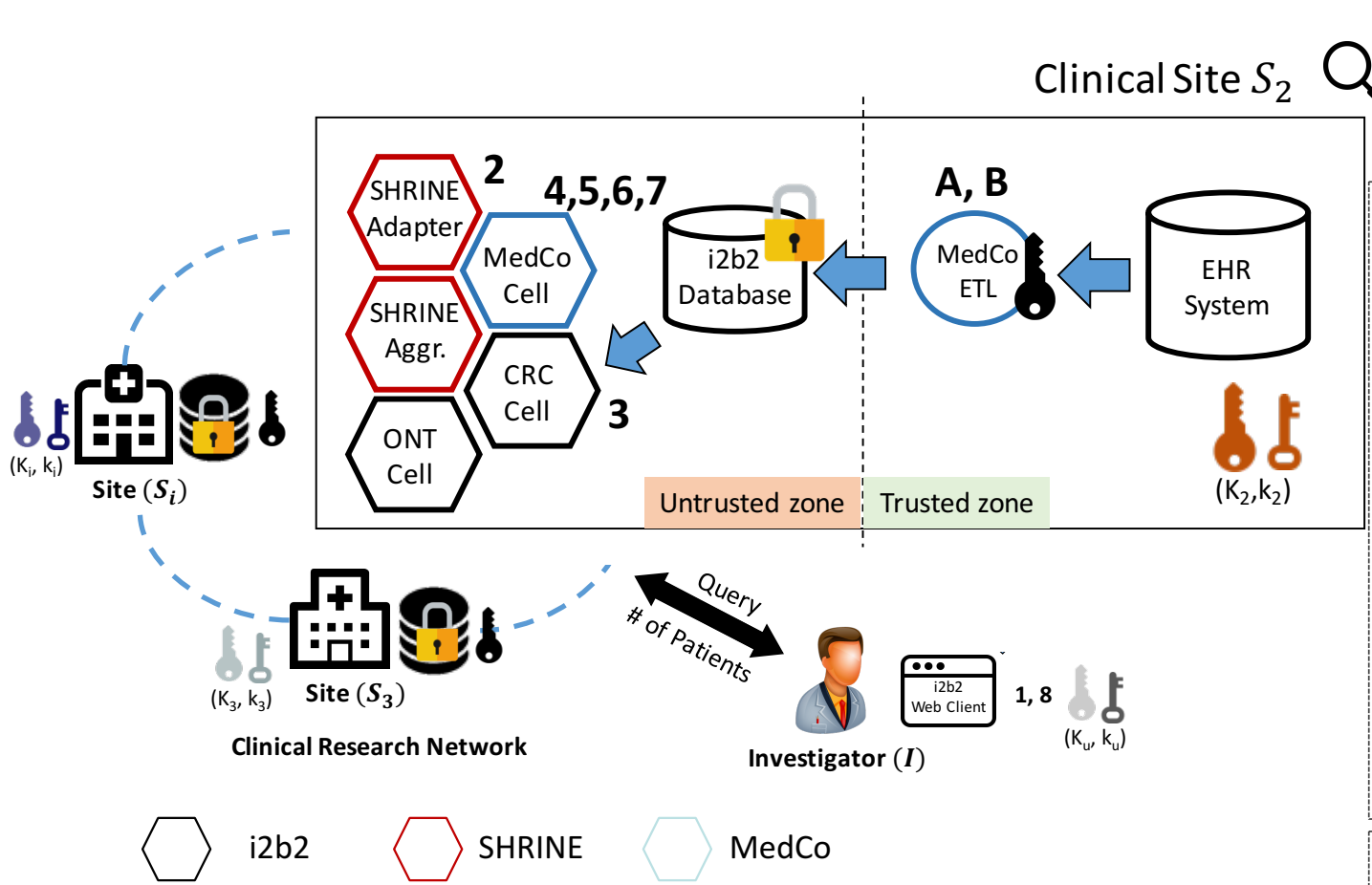
Hardware and Software Setting

- 3 servers: 2.5GHz Intel Xeon E5-2680 v3 CPUs with 12 cores
- memory: 256GB of RAM
- network: 10 Gbps link
- OS: Ubuntu



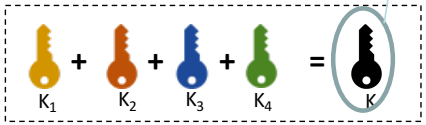
- crypto: ElGamal encryption on Ed25519 elliptic curve with 128 bit security
- database: PostgreSQL
- deployment technology: Docker

MedCo: Core Architecture & Protocol



Collective Key

Initialization phase



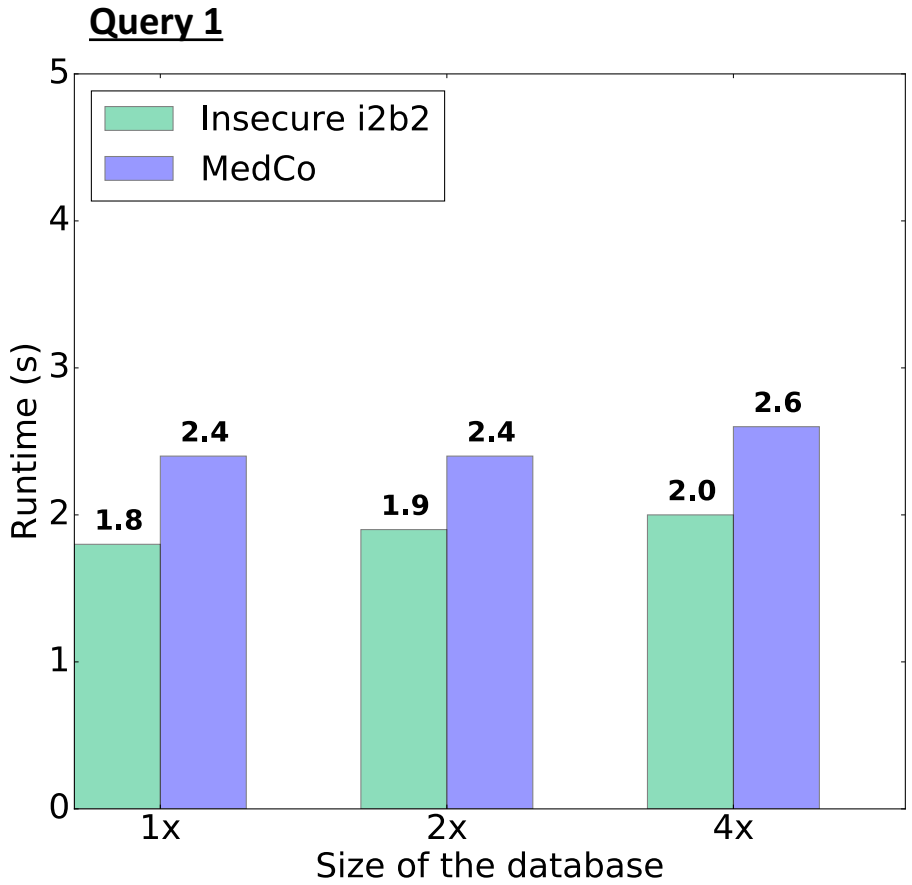
MedCo Protocol:

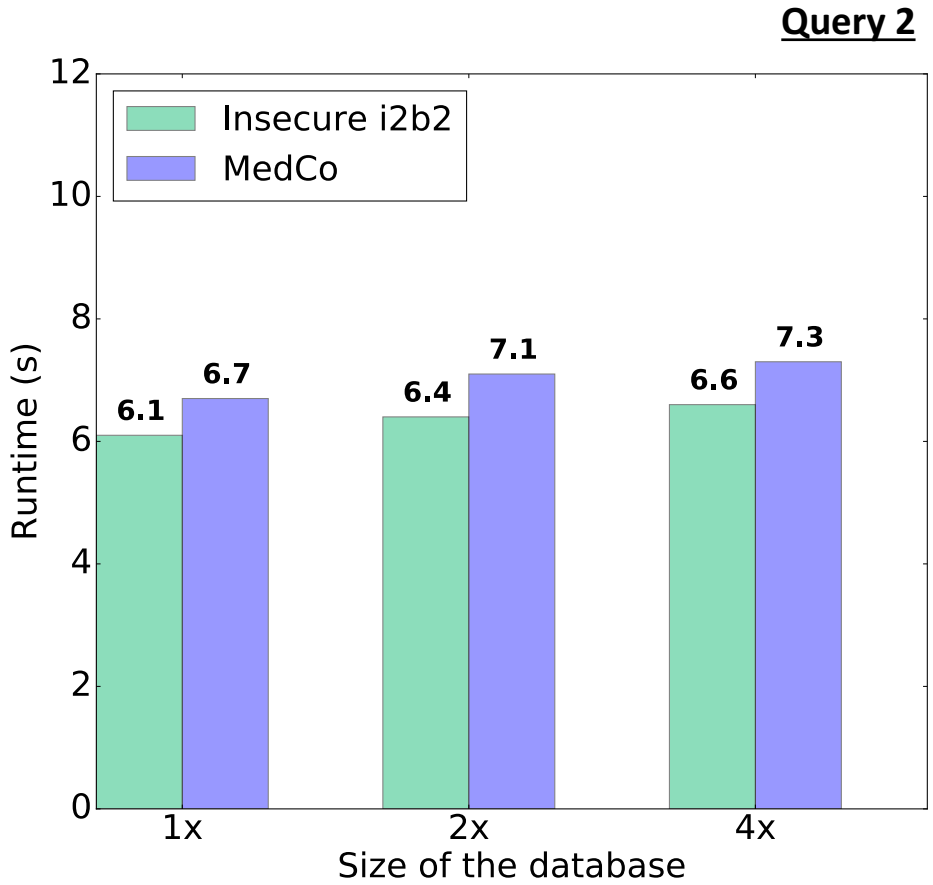
A, B) ETL & Encryption Phase

1) (user) Query Generation
2) (local) Query Analysis & (distributed) Tagging
3) (local) Query Processing
4) (local) Result Aggregation
5) (local) Result Obfuscation
6) (distributed) Results Shuffling
7) (distributed) Results Re-Encryption
8) (user) Result Decryption

ONT: Ontology
ETL: Extract, Transform, Load
CRC: Data Repository (Clinical Research Chart)
Aggr: Broadcast and Aggregator (hub)

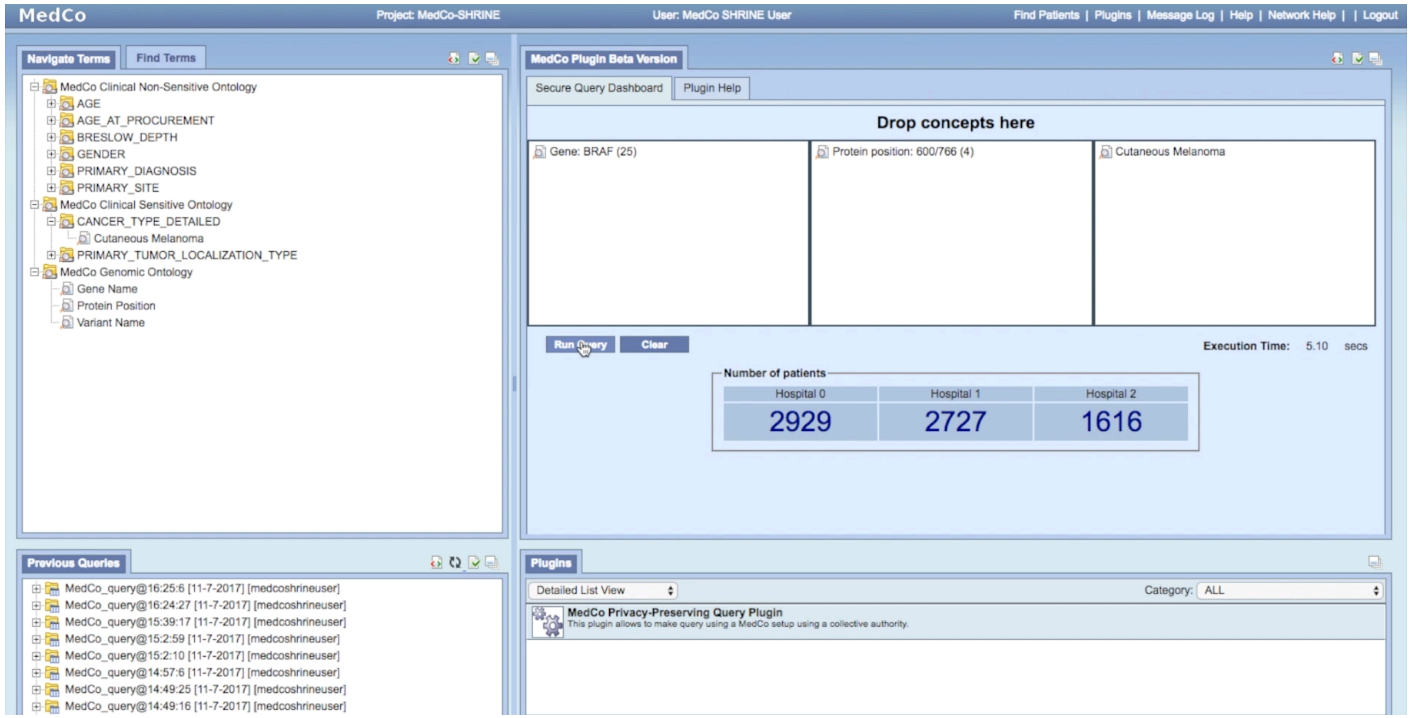
Performance Results:
Query Runtime vs. Database Size





User Experience

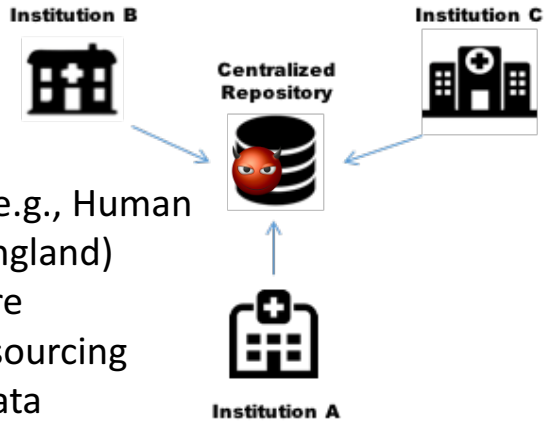
- MedCO is transparent for the investigator (through a SHRINE/i2b2 frontend plugin)



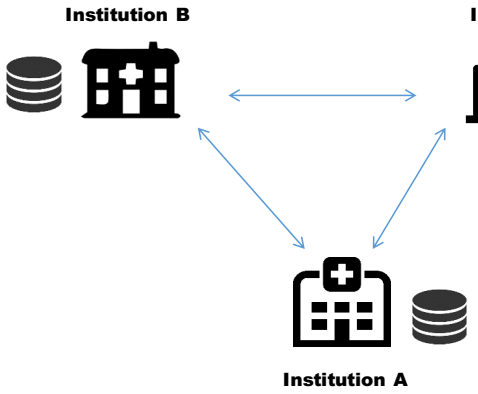
Demo of MedCo

DPPH: Data Protection in Personalized Health

- P4 (Predictive, Preventive, Personalized and Participatory) medicine
  - Revolutionize healthcare by providing better diagnoses and targeted preventive and therapeutic measures
- Challenges:
  - Scalability/Big Data
  - Efficiency and usability in data sharing
  - Mitigating privacy risks and complying with data protection
- Centralized vs Distributed Approaches:



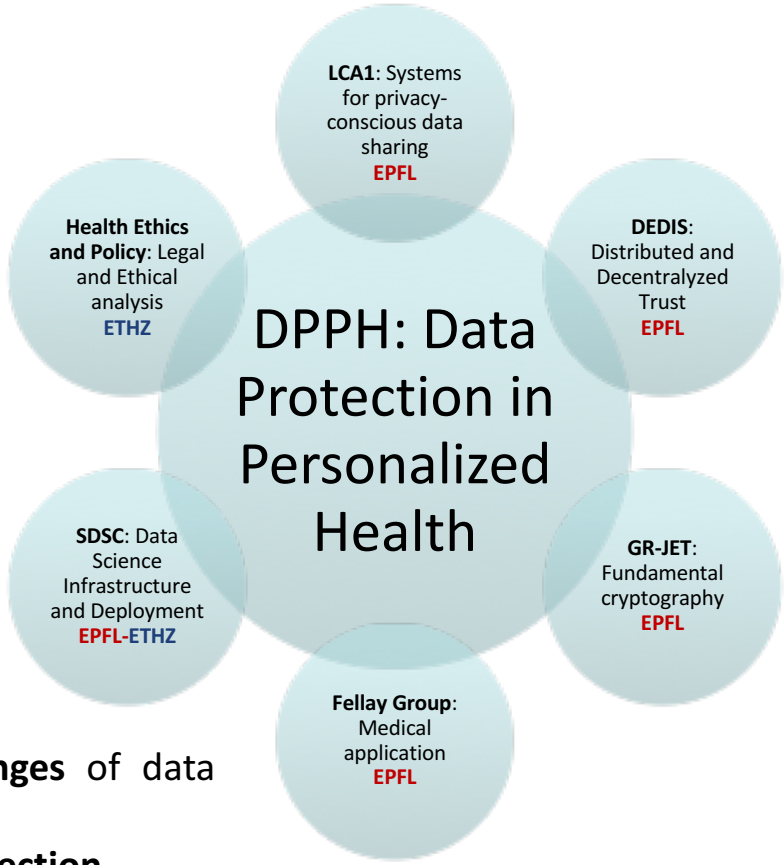
Centralized approach (e.g., Human Longevity, Genomics England)
- Single point of failure
- Impossibility of outsourcing individual patient data



Distributed approach (e.g., Beacon GA4GH, Swiss Personalized Health Network - SPHN)
- Distributed trust
- Take computation to the data

DPPH: Key Facts

- 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)
- Related event on February 15th, 2018
  - Workshop on Secure, Privacy-Conscious Data Sharing
  - http://dpph18.epfl.ch



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**

DDPH: Key Enablers

The integration of the shown tools and frameworks provides **crucial benefits for medical research**

Scalable distributed scientific computing infrastructure

Scalability and Reproducibility

Ease querying and aggregating distributed medical data

Accessibility and Usability

Secure and privacy-conscious data sharing and processing for medical data

Privacy and Accountability

Robust support for secure operations on distributed data

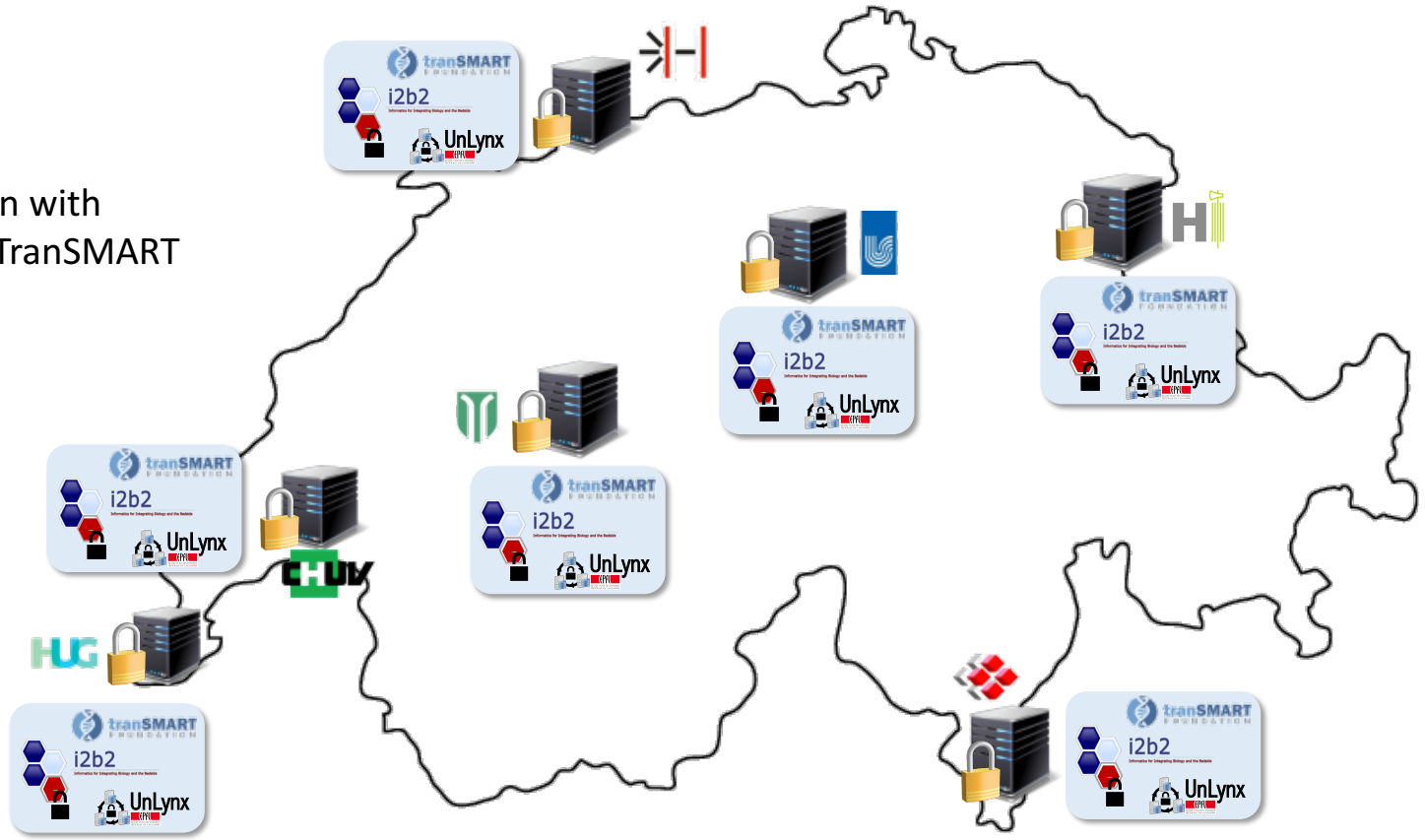
Security and Access Control

Framework for inference risks and countermeasures, and ethical analysis of distributed platforms for medical data sharing

Legal and Ethical compliance

Envisioned Nation-Wide Deployment

Integration with
I2b2 and TranSMART



Z. Huang et al: A Privacy-Preserving Solution for Compressed Storage and Selective Retrieval of Genomic Data, December 2016

JPH, S. Katzenbeisser and B. Malin Eds. Sept/Oct. 2017

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)**
- **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
- Workshop at EPFL, Switzerland, February 15, 2018

➔ Lots of material online

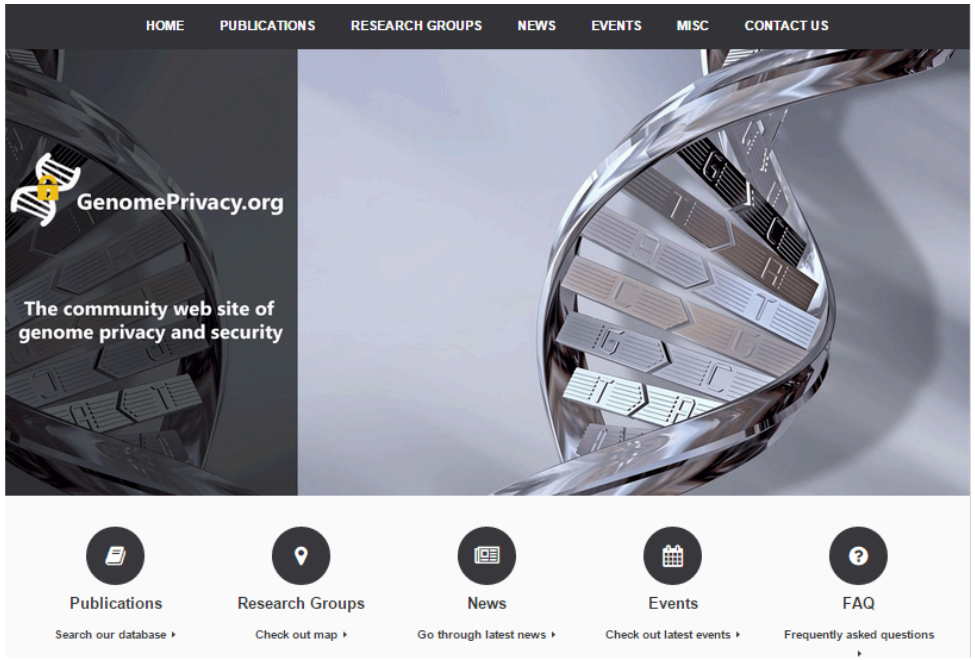
SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

iDASH

ipam

DPPH18

“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)

Conclusion

- Worldwide, medical confidentiality is **in jeopardy**
- P4 medicine requires collecting and sharing **many more data**
- Presence of **genomic data** and health-data collection with **wearable devices** will further increase the risk
- Several solutions, including **advanced cryptography**, are usable to protect genomic (and more generally medical) data
- We are working on **fully decentralized tools** (UnLynx, MedCo)
- We have **operational prototypes**, currently in deployment phase (at Lausanne University Hospital (CHUV))
- We aim at **nation-wide** deployments
- There is a tremendous need for **standardization**, especially for multi-site studies
- Our contributions to the topic:
  http://lca.epfl.ch/projects/genomic-privacy/