ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Summer School on Real-World Crypto and Privacy
Sibenik, 8 June 2017

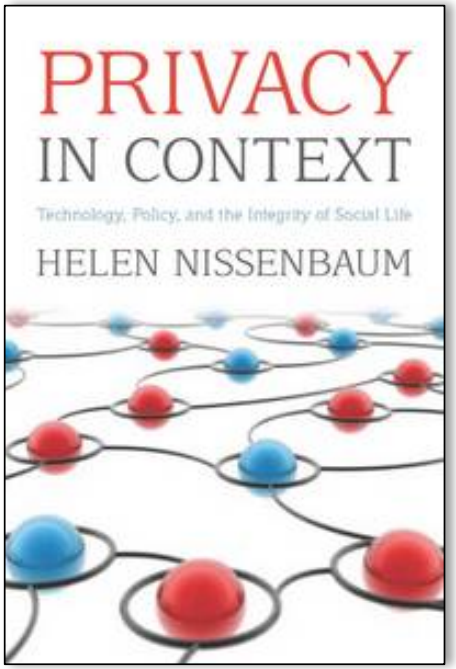
The Security and Privacy Challenges Raised by Precision Medicine

Jean-Pierre Hubaux

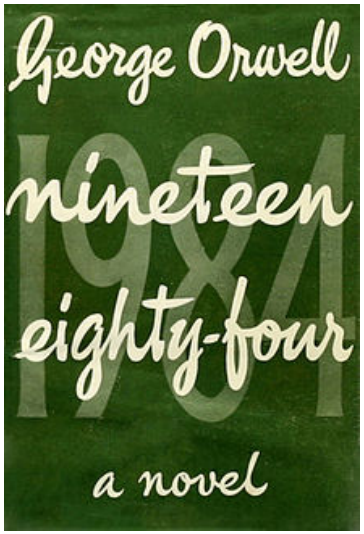
With gratitude to biomed researchers
J. Fellay, Z. Kutalik, C. Lovis, O. Michielin, V. Mooser, A. Telenti, D. Trono, P. Tsantoulis and I. Xenarios,
to CS researchers
B. Ford + team, E. Ayday, P. Egger, D. Froelicher, Z. Huang, M. Humbert, A. Juels, C. Mouchet, J.-L. Raisaro, J. Sousa, C. Troncoso and J. Troncoso-Pastoriza,
and to Sophia Genetics

Privacy: Definition

- **Privacy control** is the ability of individuals to determine when, how, and to what extent information about themselves is revealed to others.
- **Goal**: let personal data be used only in the context they have been released

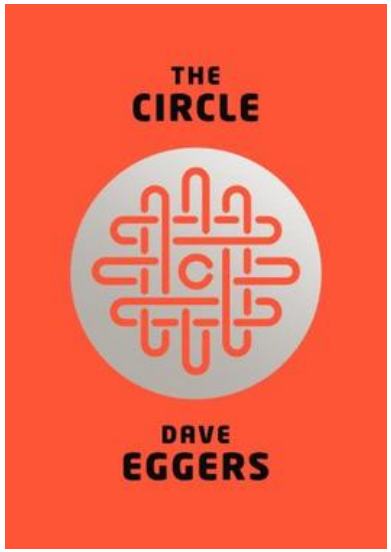


Fiction Related to Privacy



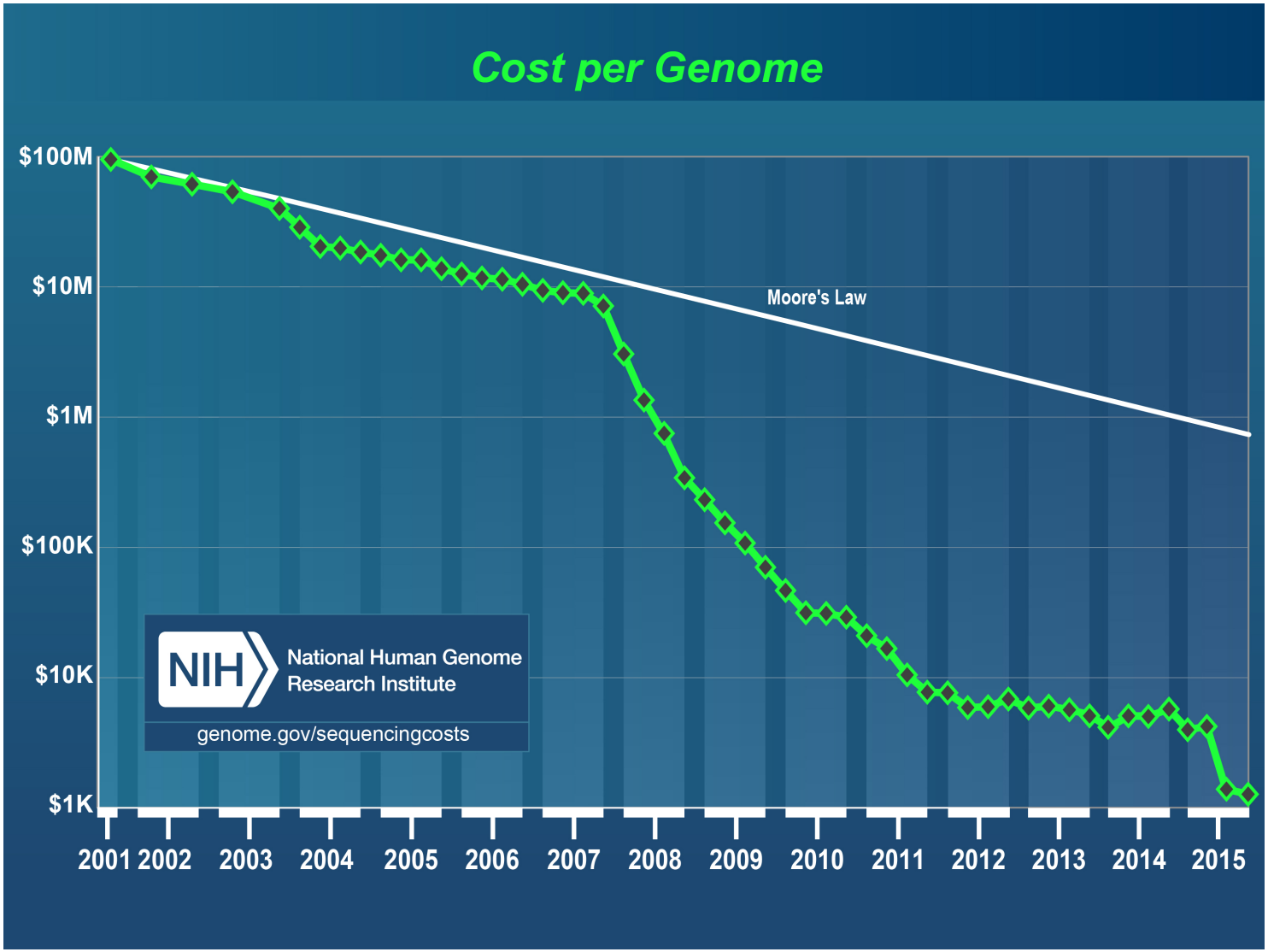
1949

The Lives of Others, 2006



2013

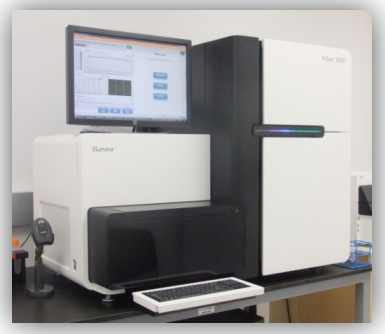
The genomic avalanche Is coming…



http://www.genome.gov/sequencingcosts/

From Blood Sample to Genome Analysis

Samples



Sequencing machine

Raw data (short reads) FastQ files

Alignment

SAM/BAM file (aligned reads)

Variant call

VCF Files

Delta with respect to the *reference genome*

ID chromosome mapping quality
Alignment Flag position Cigar string

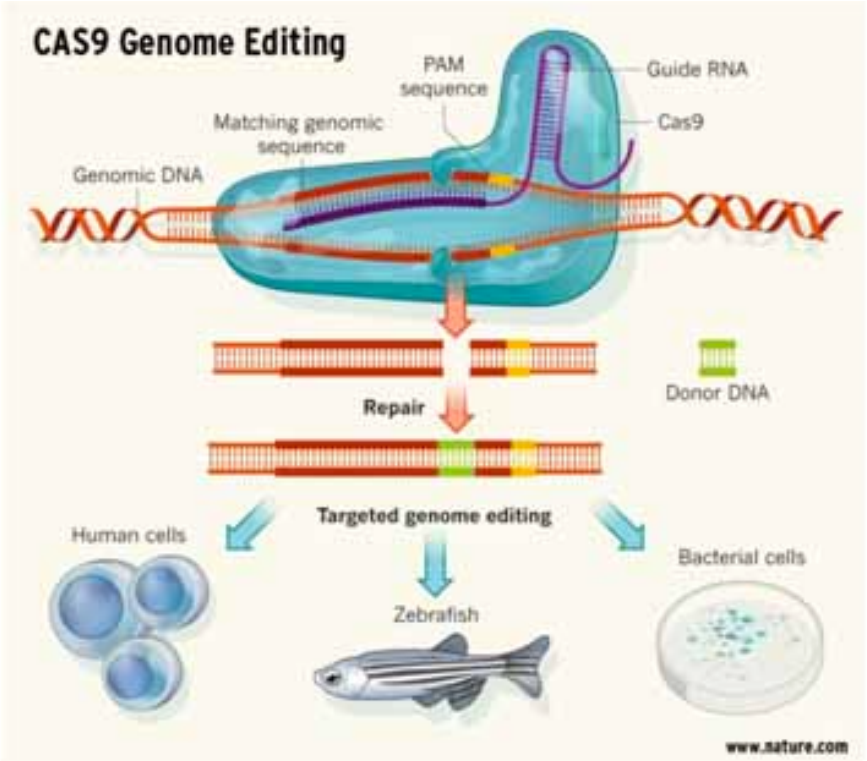
```
GZQG54D02DI67S 16   10   72065884   228   3S23M1I255M8S  *   0   0
TAGCCCCCACCCAACACACACACACGTCTTTTTTTCTGAATCATTTGAGTGTACCATACATAC
GTCATGGCCCTCAGTGTATATTTCCTAGGAATTGGGACATTCTCTTACATAACCACAGTGCAGT
TAAGGGTTCCAGGTATTCACATTGATAGAATGTGTATCTAACCCATGGCCCACAGTTCAGTGT
GGTCAGGTGCCCTAACATCCTGTACAGCACTTCCCCCTCTAGTACCAGAGCCAGTCTGGAAT
CAAGGATTGCATTTACTTCTAACTCTTTGGCCCCTCTG     E=7/////200088;@;;;@EEIID?,?
2666644IHHHHIIH<<<HIIIIIIIIIIIIIIIIII;;555CEIEEIIIII???
HHIIIIIIIHHHIIIIIIIIIIIHDCCIIIIIIIIIIIIIIIIHHHCCCCIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHHIIIIHHHIIIDCCIIIIIIIIIIIIIIHH
HIIIIIIIIIIIIIII>CC;;;;;;HIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHHIIIIIHHHIII;;;IIAAAAIIII     AS:i:267     XS:i:0  XF:i:0
XE:i:6  XN:i:0
```


3 billion letter pairs, with high *coverage*, to take into account:

- **Sequencing errors**
- **Possible mutations**

Genome Editing (CRISPR-CAS9)

- Potential to alter the **human** genome
- Strong potential for treatment of (human) genetic diseases
- Moratorium pronounced in December 2015 for edition of inheritable parts of the human genome
- Used at least once on monkeys in China



CRISPR: Clustered regularly interspaced short palindromic repeats
CAS9 is a protein

Medical Use of Genetics

- Genetic disease risk tests help early diagnosis of serious diseases
- Pharmacogenomics ➔ personalized medicine

The Genomic Era

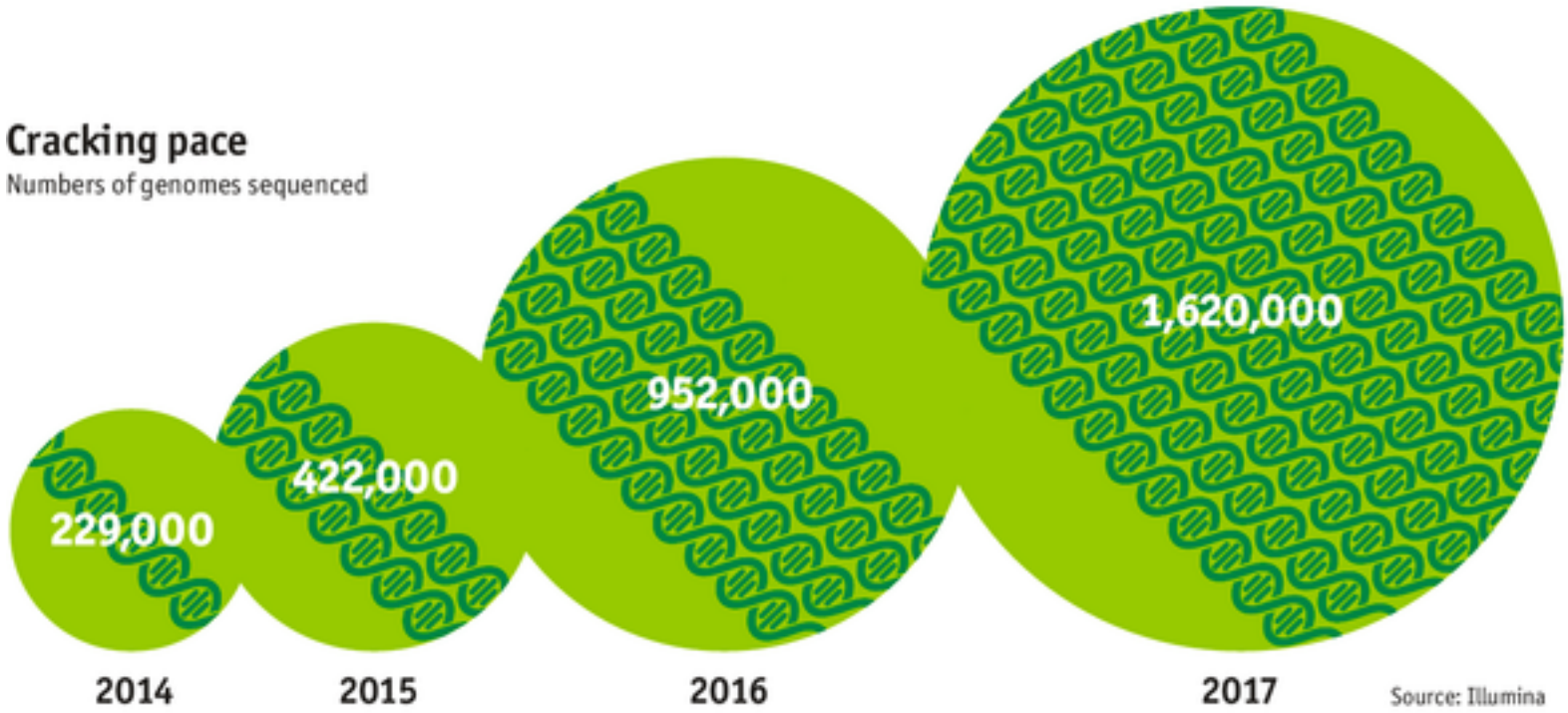


Figure from The Economist


MIT Technology Review
VOL. 118 NO. 3 MAY/JUNE 2015 $6.99
Feature p. 48
HP Tries to Reinvent the Computer
Business Report p. 63
Persuasion
Review p. 72
The Problem with Fake Meat
WE CAN NOW ENGINEER THE HUMAN RACE
p26

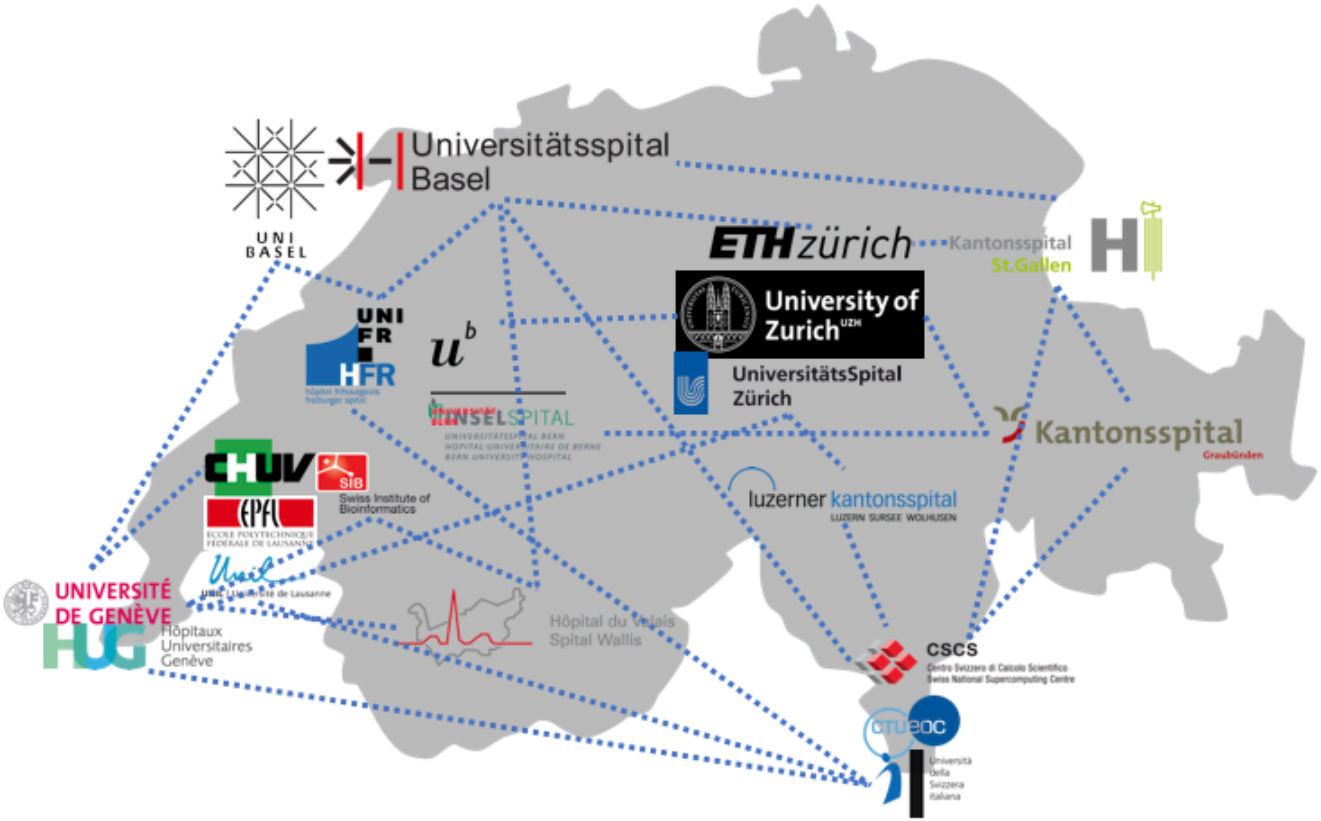
Governmental Initiatives on Genomics

- August 2014: Prime Minister Cameron Project – Genomics England
  → 100’000 citizens

- January 2015: President Obama’s Precision Medicine Initiative
  → 1,000,000+ citizens

Swiss Personalized Health Network (SPHN)

- National initiative launched by the Swiss Federal Government (2017-2020+)
- Goal: create a national infrastructure enabling the sharing across Switzerland of patient data for research and clinical care



Industry Initiatives

- IT giants start proposing genome-related services
  - Google Genomics (API to store, process, explore, and share DNA data)
  - IBM Research (computational genomics)
  - Microsoft Research (genomic research in collaboration with Sanger Center)
  - Apple (the ResearchKit program)
  - Amazon


- Global Alliance for Genomics & Health

  - Definition of a common framework for effective, responsible and **secure** sharing of genomic and clinical data
  - Security Working Group: security infrastructure policy and technology
    http://genomicsandhealth.org/working-groups/security-working-group

Privacy-Conscious Exchange of Medical Data: Analogy

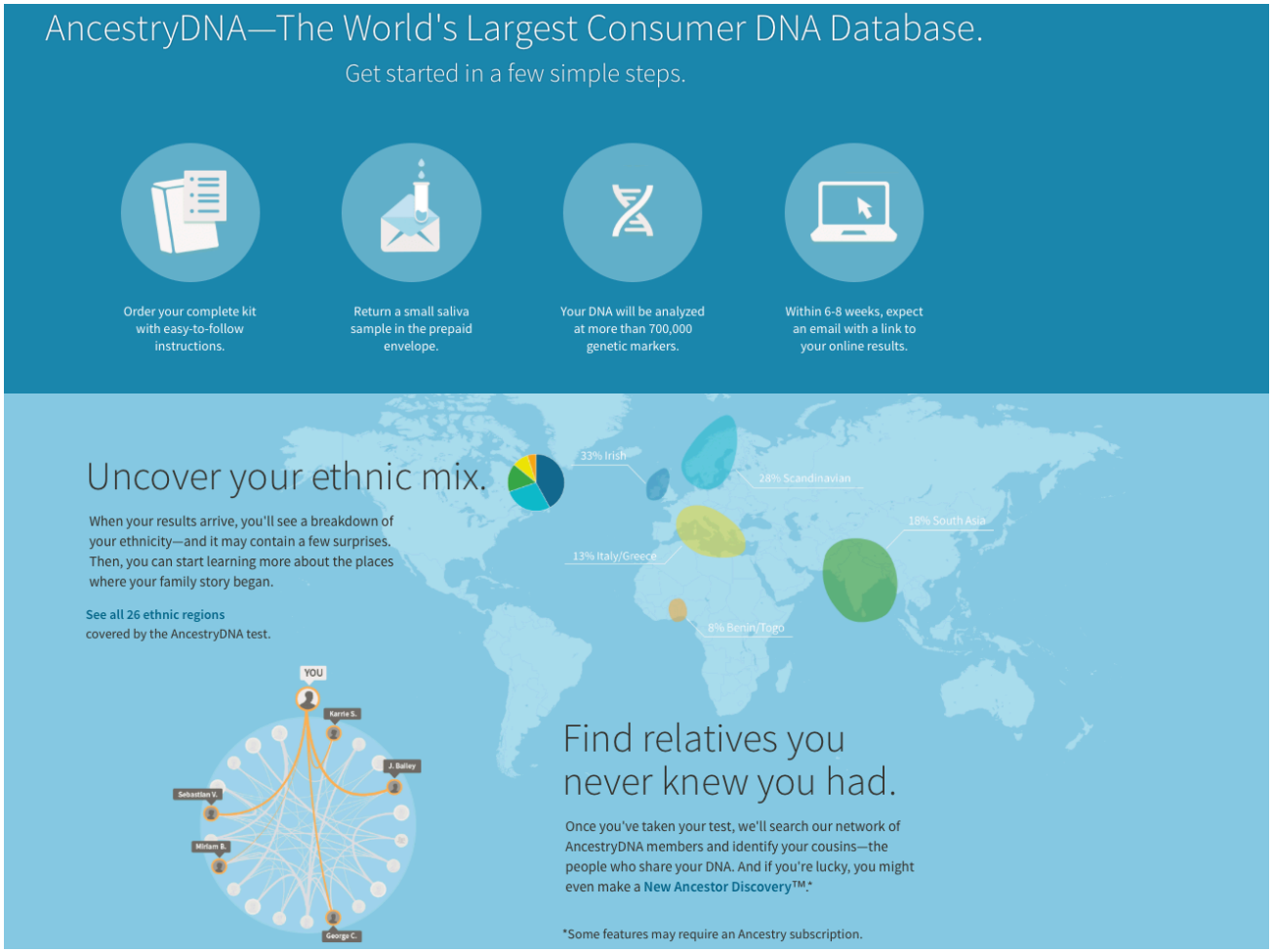
→ Exchange of data related to personalized medicine

→ World-Wide Web protocols

→ Internet protocols

Direct-to-Consumer Genomics (1/2)

- Ancestry.com (1 million+ customers)



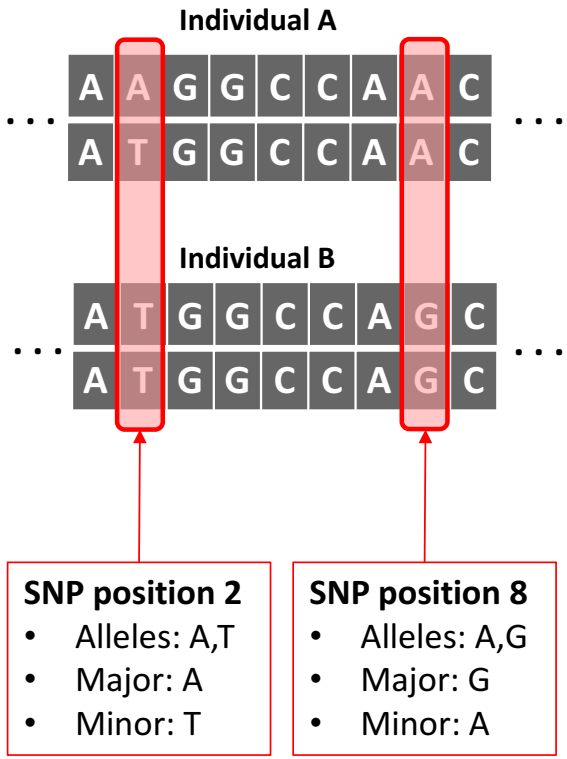
Direct-to-Consumer Genomics (2/2)

- 23andMe.com
  (1 million+ 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%

Most common genetic variation: Single Nucleotide Polymorphism (SNP)

- Occurs when, at a specific position, at least a single nucleotide (A,C,G, or T) differs between members of the same species in more than 1% of the population
- Potential nucleotides for a SNP are called **alleles**
- 2 different alleles can be observed for each SNP:
  - **Major** allele (M)
  - M**inor** allele (m)
- Every genome carries 2 alleles at each SNP position
  A SNP can be either:
  - **Homozygous minor** [m,m]
  - **Heterozygous** [m,M] or [M,m]
  - **Homozygous major** [M,M]



Alice and Bob:
The Long-Awaited Happy End

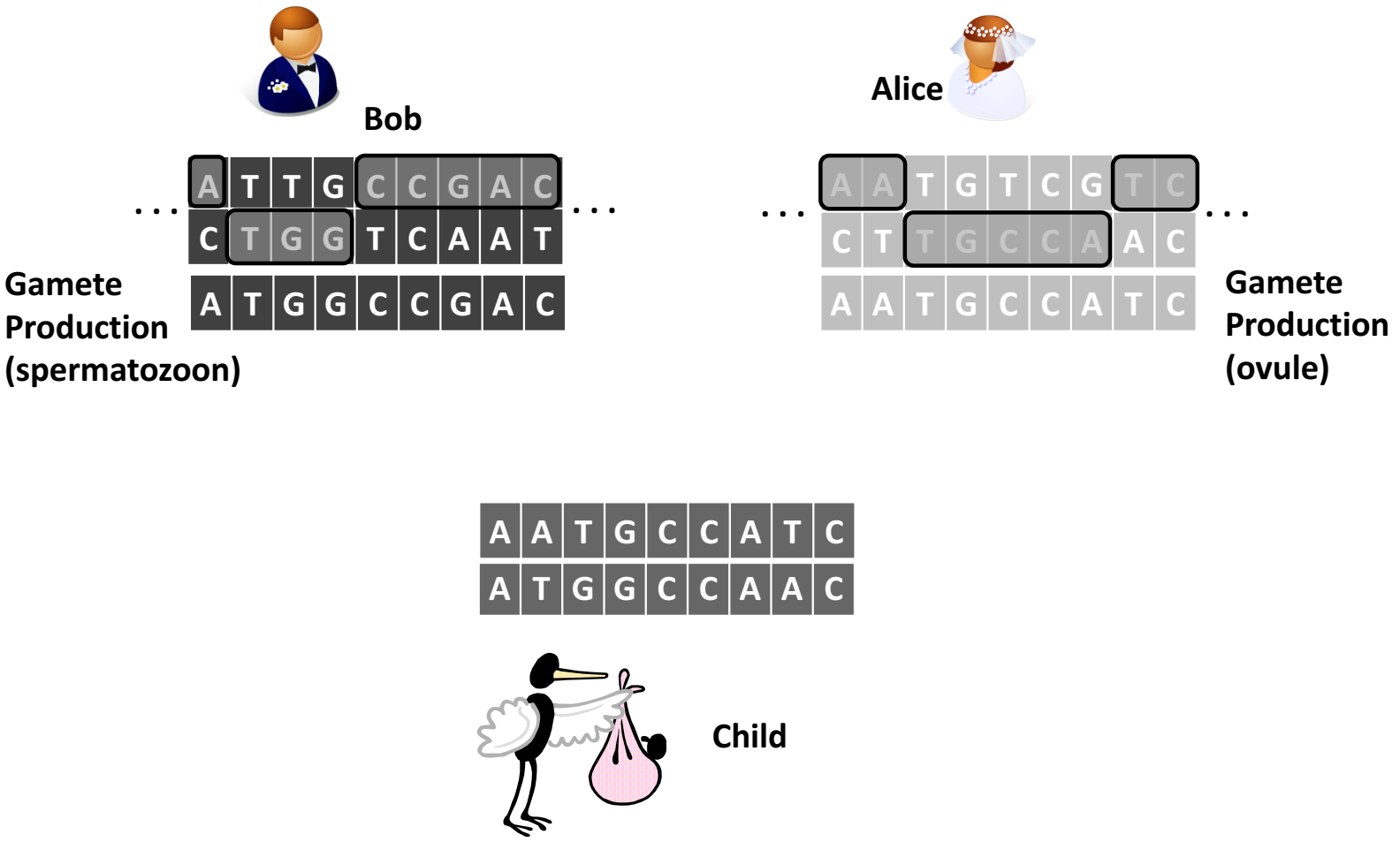
After having extensively authenticated each other,

after having exchanged thousands of highly private messages,

after having established numerous secure channels between each other,

after years of intense but platonic relationship, finally, finally…❤️

… Alice and Bob got closer to each other



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) +

“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

Hacking of Anthem Insurance

Anthem Hacking Points to Security Vulnerability of Health Care Industry

By REED ABELSON and MATTHEW GOLDSTEIN FEB. 5, 2015

The New York Times

An Anthem Health Insurance facility in Indianapolis. Hackers gained access to about 80 million company records, and some fear the stolen data will be used for identity theft. Aaron P. Bernstein/Getty Images

- Anthem: one of US largest health insurers
- 60 to 80 million *unencrypted* records stolen in the hack (revealed in February 2015)
- Contain social security numbers, birthdays, addresses, email and employment information and income data for customers and employees, including its own chief executive

US Healthcare “Wall of Shame”

On average, one breach is declared **every day**, each affecting 500+ people

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (since 2009)

Breaches Affecting 500 or More Individuals

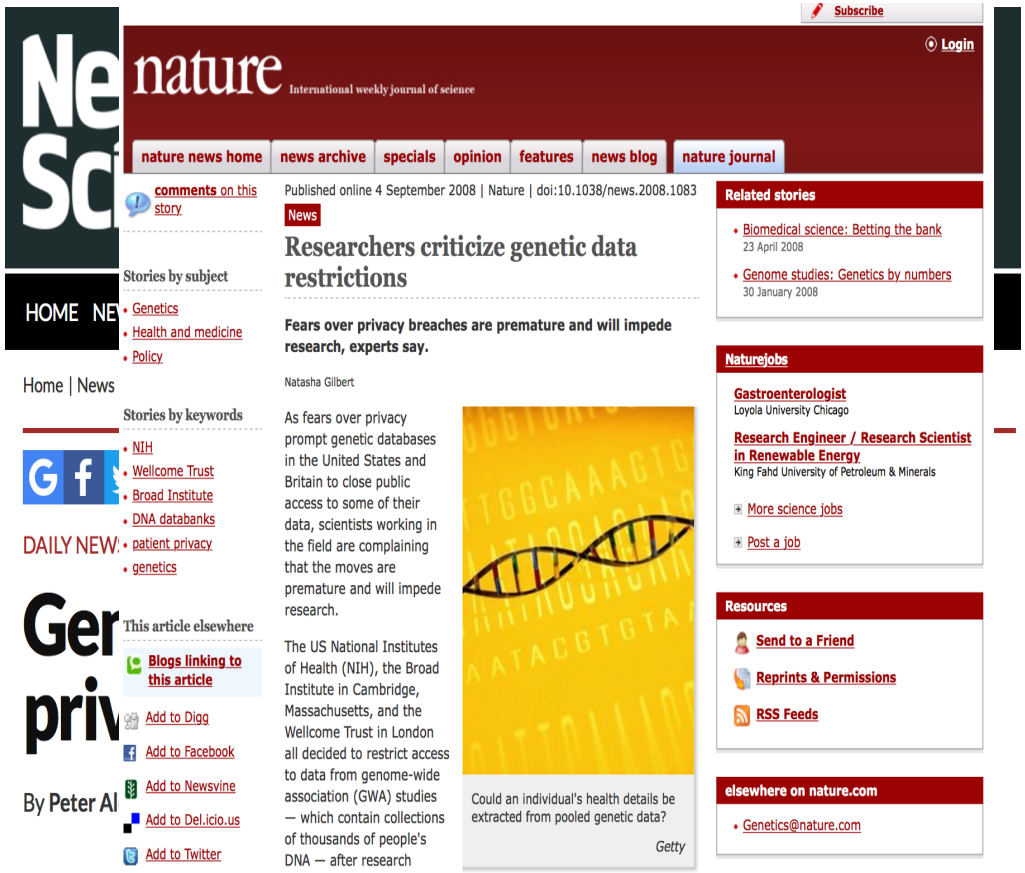
As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals. These breaches are now posted in a new, more accessible format that allows users to search and sort the posted breaches. Additionally, this new format includes brief summaries of the breach cases that OCR has investigated and closed, as well as the names of private practice providers who have reported breaches of unsecured protected health information to the Secretary. The following breaches have been reported to the Secretary:

Show Advanced Options

Breach Report Results

Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information
New York City Health and Hospitals Corporation - Coney Island Hospital	NY	Healthcare Provider	3494	05/09/2017	Unauthorized Access/Disclosure	Other, Paper/Films
Clinton County Board of Developmental Disabilities	OH	Healthcare Provider	1243	05/05/2017	Hacking/IT Incident	Network Server
Mecklenburg County, North Carolina	NC	Healthcare Provider	2000	05/04/2017	Unauthorized Access/Disclosure	Other Portable Electronic Device
LSU Healthcare Network	LA	Healthcare Provider	2200	05/04/2017	Theft	Other Portable Electronic Device
Capital Nephrology	MD	Healthcare Provider	4000	05/02/2017	Hacking/IT Incident	Electronic Medical Record, Network Server
Nova Southeastern University	FL	Healthcare Provider	1086	05/02/2017	Theft	Other Portable Electronic Device
Michagan Facial Aesthetic Surgeons d/b/a University Physician Group	MI	Healthcare Provider	3467	04/28/2017	Theft	Laptop

Another Major Concern: Re-identification Attacks against Genomic Databases



Subscribe

Login

nature International weekly journal of science

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject
- Genetics
- Health and medicine
- Policy

Home | News

Stories by keywords
- NIH
- Wellcome Trust
- Broad Institute
- DNA databanks
- patient privacy
- genetics

This article elsewhere
- Blogs linking to this article
- Add to Digg
- Add to Facebook
- Add to Newsvine
- Add to Del.icio.us
- Add to Twitter

Published online 4 September 2008 | Nature | doi:10.1038/news.2008.1083

News

Researchers criticize genetic data restrictions

Fears over privacy breaches are premature and will impede research, experts say.

Natasha Gilbert

As fears over privacy prompt genetic databases in the United States and Britain to close public access to some of their data, scientists working in the field are complaining that the moves are premature and will impede research.

The US National Institutes of Health (NIH), the Broad Institute in Cambridge, Massachusetts, and the Wellcome Trust in London all decided to restrict access to data from genome-wide association (GWA) studies — which contain collections of thousands of people's DNA — after research

Could an individual's health details be extracted from pooled genetic data?

Getty

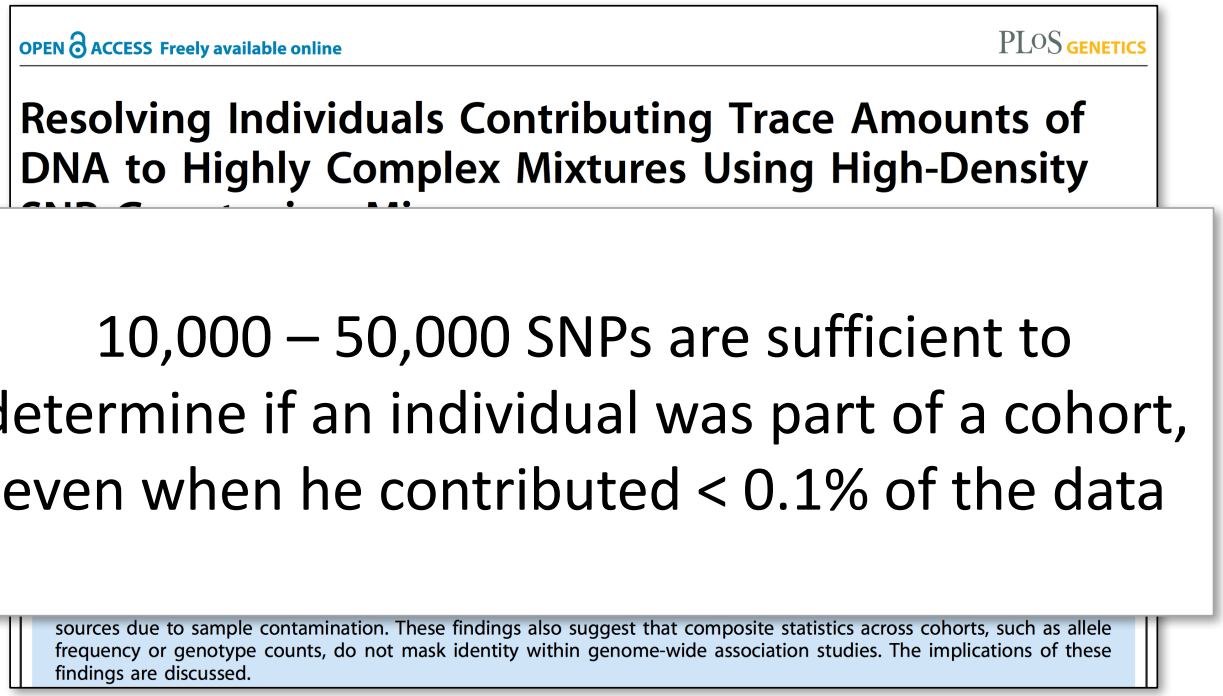
Related stories
- Biomedical science: Betting the bank 23 April 2008
- Genome studies: Genetics by numbers 30 January 2008

Naturejobs
- Gastroenterologist, Loyola University Chicago
- Research Engineer / Research Scientist in Renewable Energy, King Fahd University of Petroleum & Minerals
- More science jobs
- Post a job

Resources
- Send to a Friend
- Reprints & Permissions
- RSS Feeds

elsewhere on nature.com
- Genetics@nature.com

Re-identification Attacks on Genomic Data

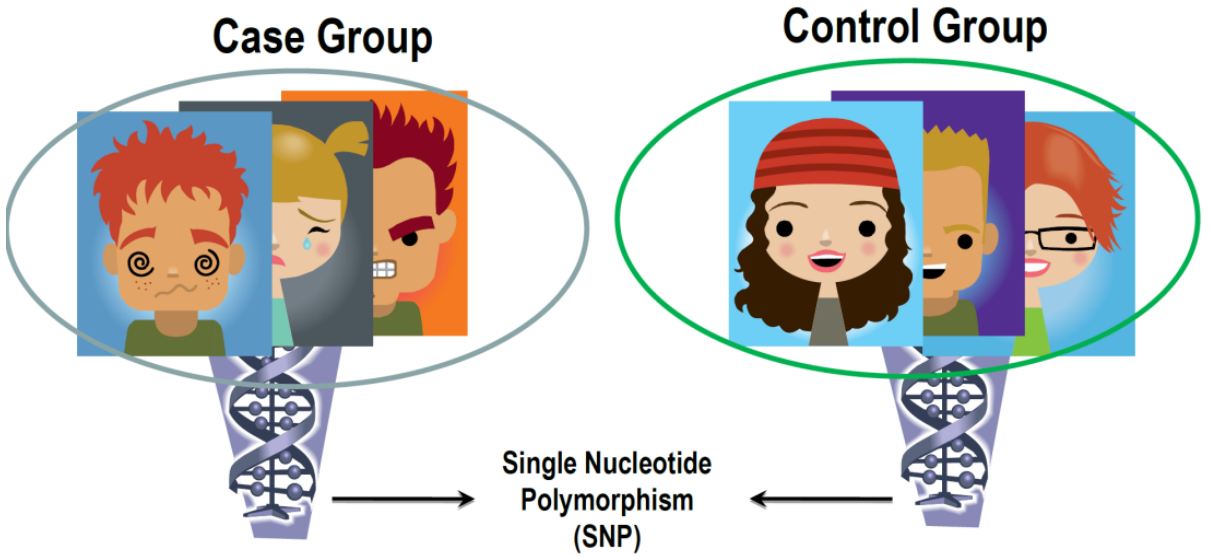


Many other subsequent studies extended the range of vulnerabilities for summary statistics:

[Jacobs et al. *Nature Genet.* '09], [Vissecher and Hill *PLoS Genet.* '09], [Sankararaman et al. *Nature Genet.* '09], [Wang et al. *CCS*'09], [Clayton *Biostatistics* '10], [Im et al. *Am. J. Hum. Genet.* '12], …

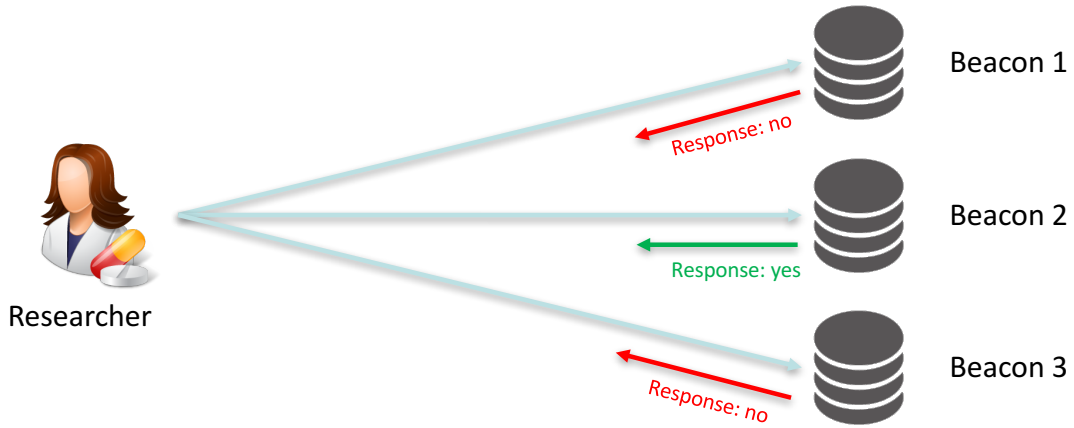
Homer Attack

- Adversary has access to a known participant's genome
- Goal: determine if the target individual is in the case group
- Uses simple correlation in the genome (linkage disequilibrium)
- Attack later improved by Wang et al.



N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics, 4, Aug. 2008.

GA4GH Beacon Project



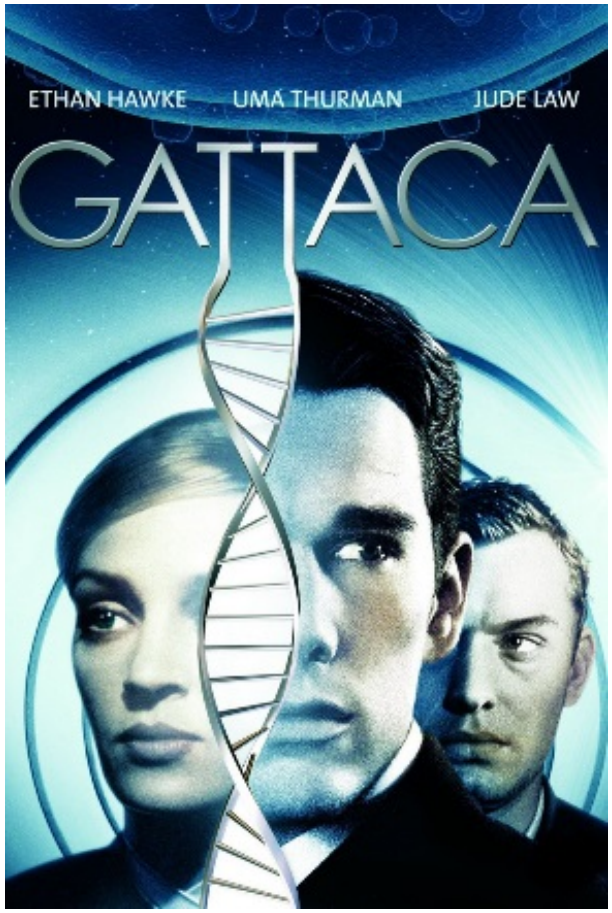
Main features:

- Enables researchers to quickly query multiple database to find the sample they need
- Encourages cross-border collaboration among researchers
- Provides only minimal responses back in order to mitigate privacy concerns

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** (hence, of requirements): healthcare, medical research, forensics, ancestry



1997

Canonical Misconception about Genome Privacy and Security

Genome privacy is hopeless, because all of us leave biological cells (hair, skin, droplets of saliva,…) wherever we go

- Those cells can be collected and used for DNA sequencing
- Hence trying to secure genomes is a lost battle
- What is wrong with this reasoning?
- Collecting human biological samples and sequencing them is expensive, illegal, prone to mistakes, and non-scalable! (even if sequencing techniques keep improving)
- The medical community (research and healthcare) **should not be** the (indirect) accomplice of massive leaks of sensitive data

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,…)

Possible Solutions

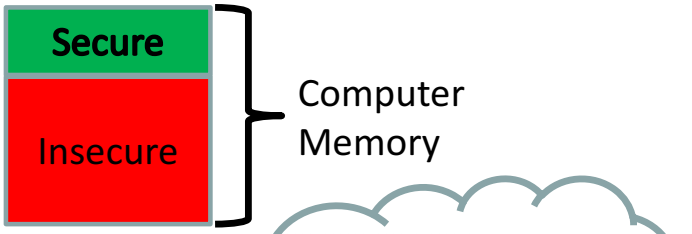
- Centralized bunker (“Fort Knox”)

- Hardware-based solutions (Intel SGX & Co)

- Cloud provider (Amazon Cloud, MS Azur,…)

- Software-based, **decentralized**, **open-source**, **provable** secure solutions, with **data staying at the hospitals**

Hardware-Based Solution: Trusted Hardware



- **Secure memory region**
- **Nothing outside the region can inspect the memory, not even the cloud operating system!**

Guaranteed by the CPU
Trusted Hardware
Example: Intel SGX

Encrypted sensitive data: E(X)

Output: F(X, Y)

Output: F(X, Y)

Encrypted sensitive data: E(Y)

Therefore, as long as the decryption and computation of sensitive data happens inside the secure memory, no one can steal the data, not even the cloud administrator.

E(X) stands for encryption of X
F(X, Y) is a computation F on inputs X and Y

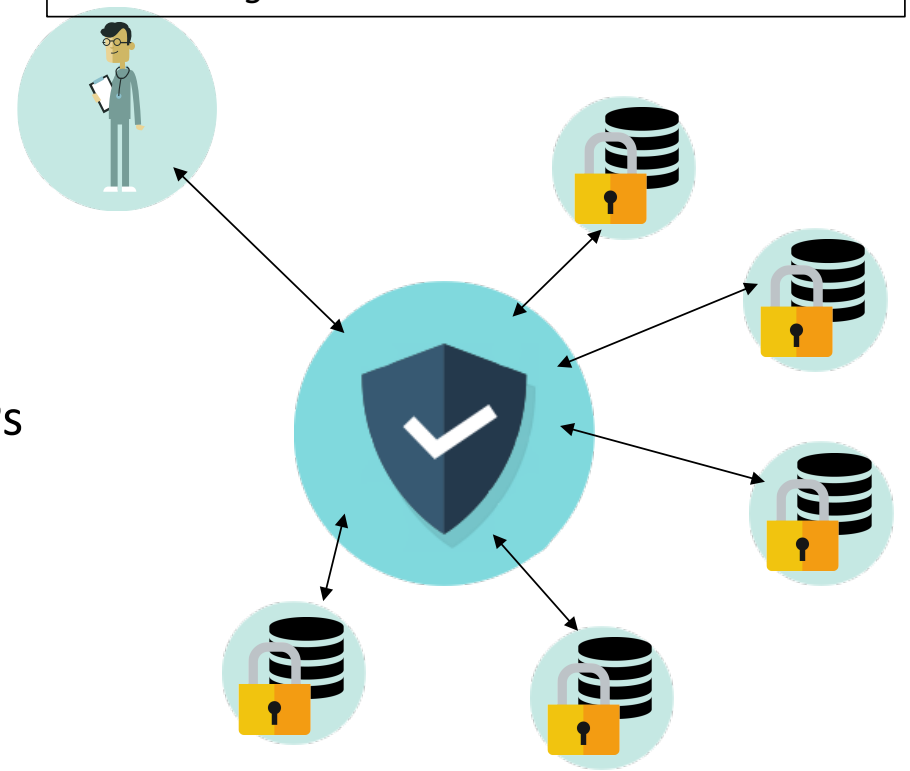
Drawbacks: - you need to trust the vendor
- side-channel attacks

Software-based, **decentralized**, **open-source**, **provable** secure solutions, with **data staying at the hospitals**:

UnLynx

Problem Statement

SELECT *AVG*(*cholesterol_rate*)
FROM DP_1, …, DP_n
WHERE *age in [40:50]* AND *ethnicity = Caucasian*
GROUP BY *gender*



Functionality:

- Enable queriers to query a set of distributed databases

Requirements:

- Confidentiality of data provided by Data Providers (DP)
- Privacy of individuals storing their data in DPs
- No single point of failure
- Computational correctness

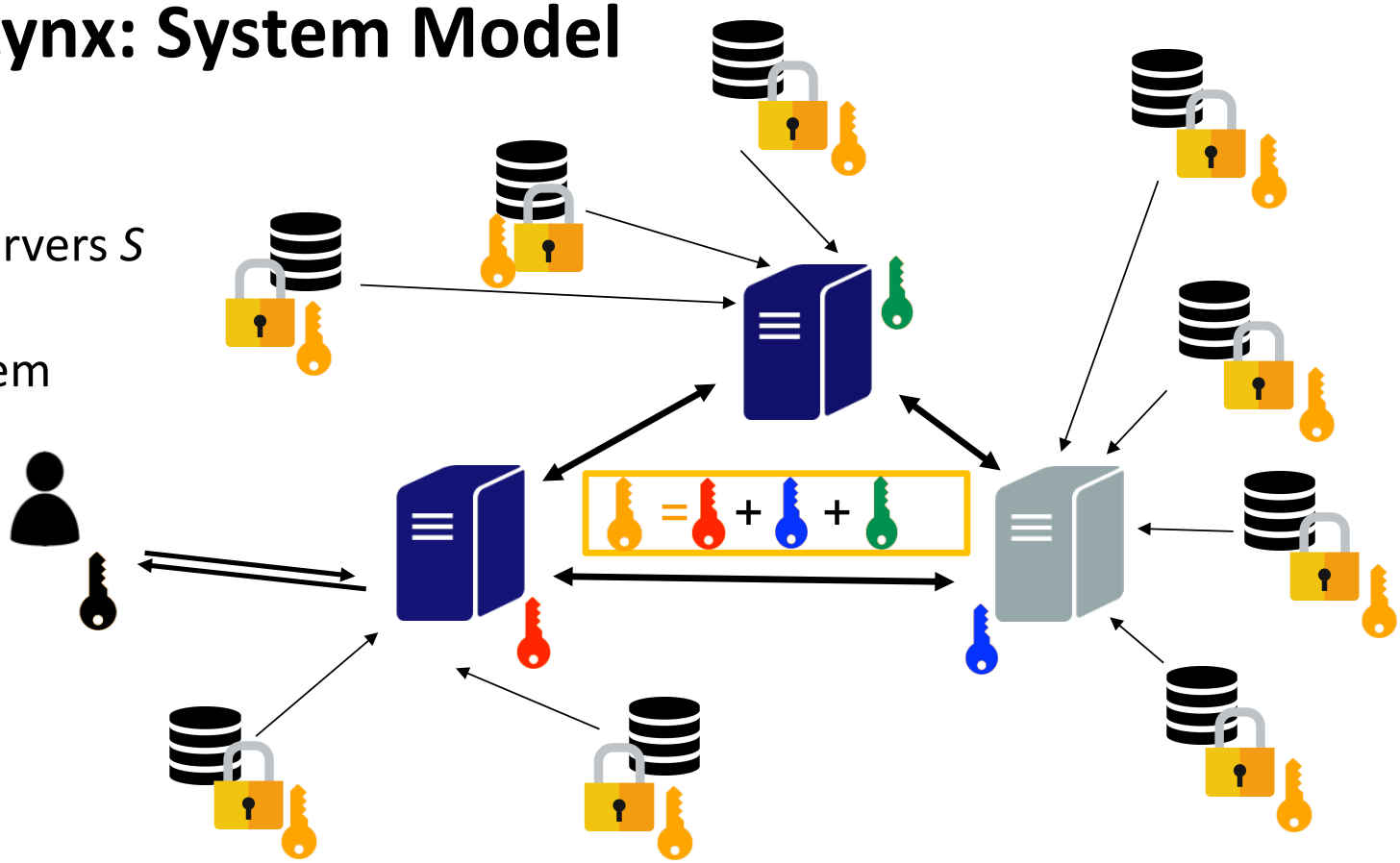
Threat model:

- Queriers and computation entity can be malicious
- DPs are honest-but-curious

UnLynx: System Model

Involved parties:

- Collective authority of *m* servers *S*
- *n* Data providers *DPs*
- Clients *Q* querying the system



At initialization, Data Providers encrypt their (sensitive) data with the public key (🔑) formed by the 3 servers.

→ secure as long as at least one of the servers is honest.

UnLynx: Security guarantees

Data are encrypted during the whole process.

Misbehaving entity can be identified and expelled.

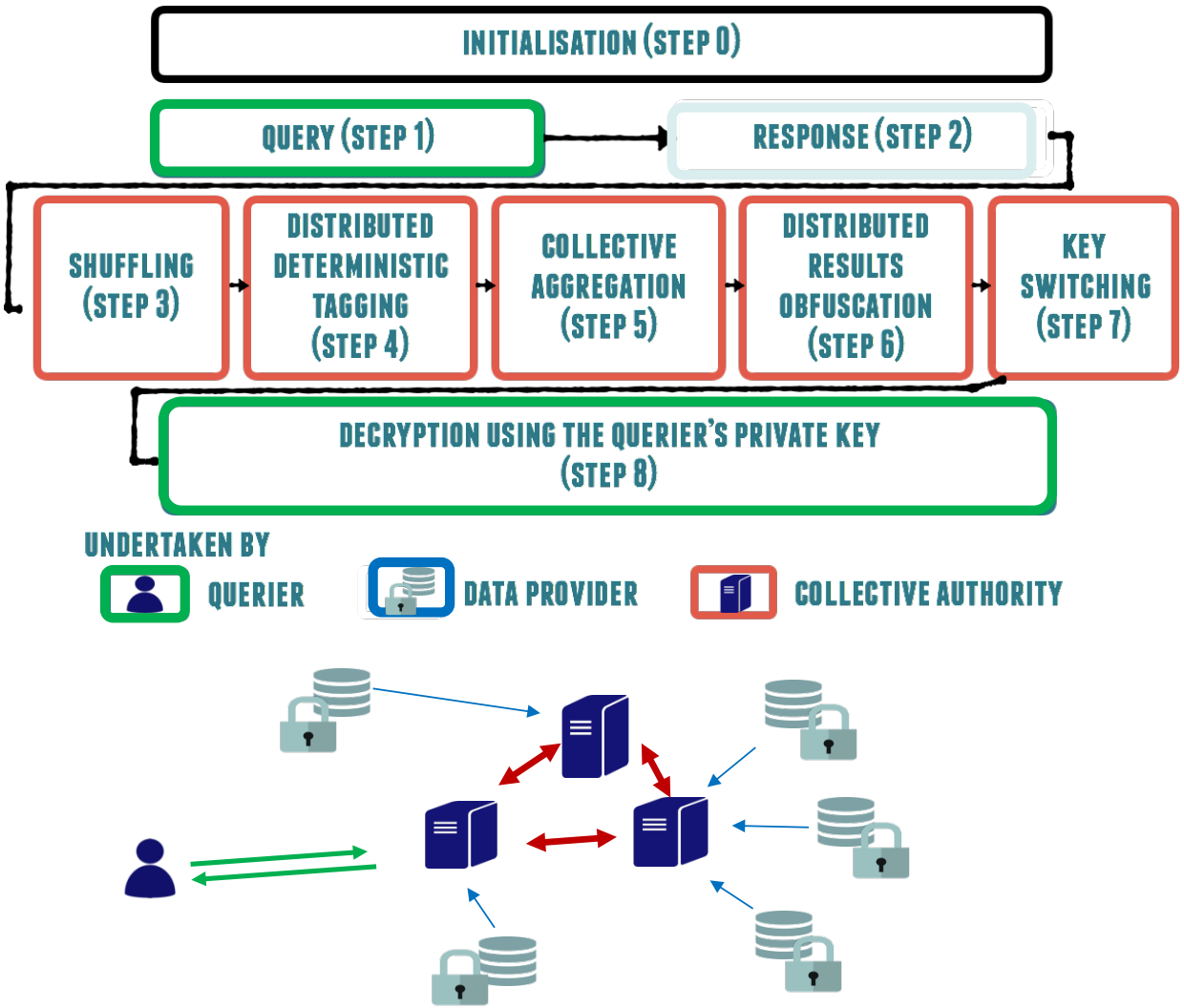
Data are shuffled to break the link btw. DP and data.

Correctness of each computation can be verified with Zero-Knowledge Proofs (proof that the computation is correct without disclosing the secret values).

Oblivious noise addition on query results ensures differential privacy.

As long as one of the servers is honest, all the other properties are guaranteed.

UnLynx - Workflow



Shuffling: break link between data and data providers

Distributed Deterministic Tagging: permits to group/filter the responses

Collective Aggregation: aggregation of all responses

Distributed Results Obfuscation: addition of noise to query results in order to ensure differential privacy

Key Switching: transform the data encryption from the collective authority public key to the researcher key without decrypting

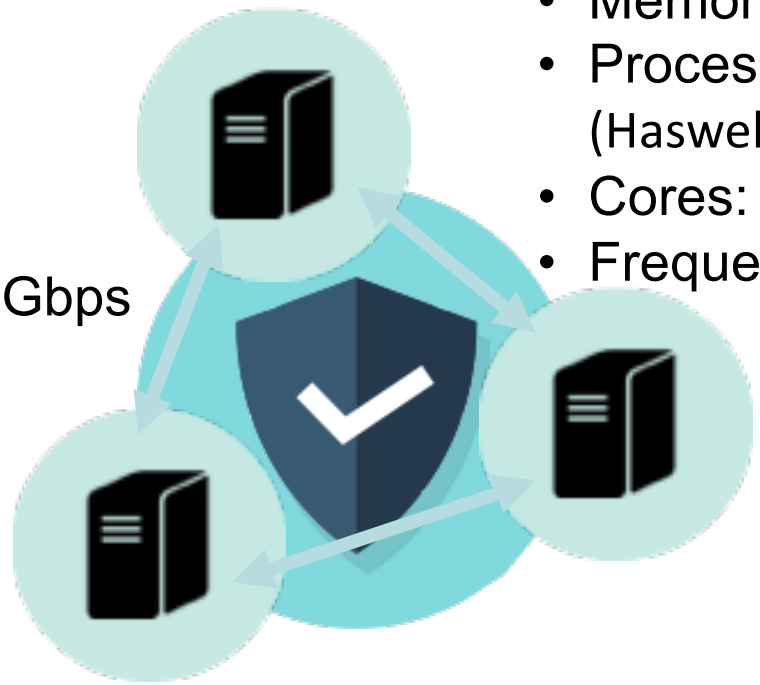
Performance Evaluation

Server:

- Memory: 256GB RAM
- Processor: Intel Xeon E5-2680 v3 (Haswell)
- Cores: 24 (with 48 threads)
- Frequency: 2.5GHz

Network:

- Bandwidth: 1Gbps
- Delay: 10ms



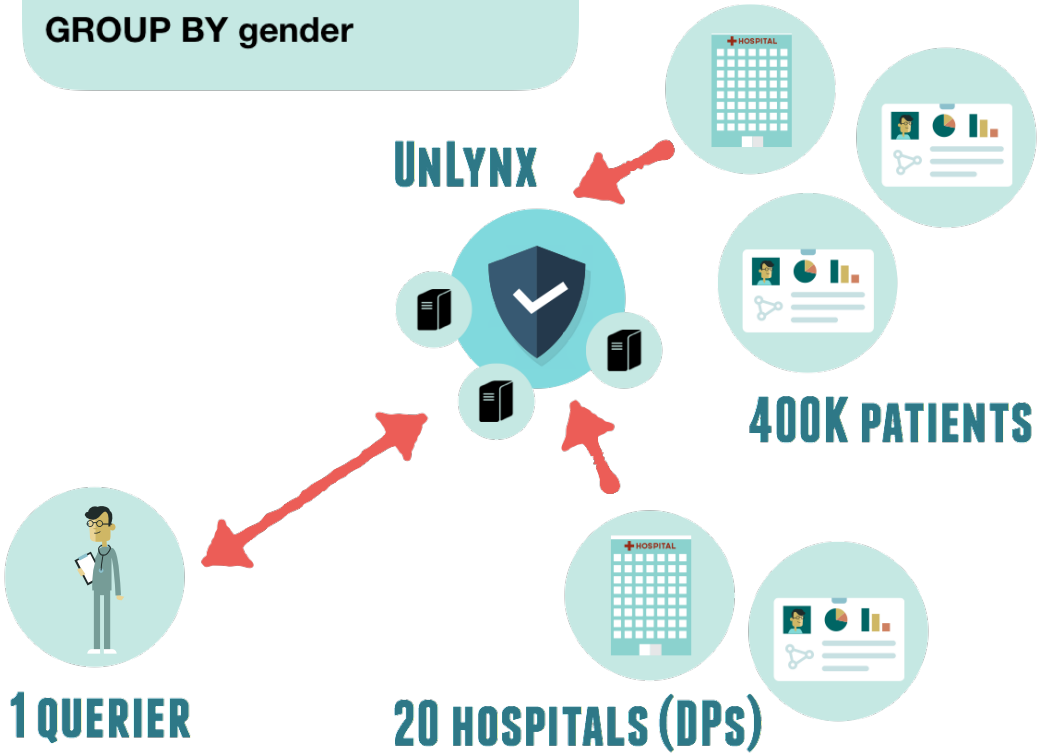
Crypto

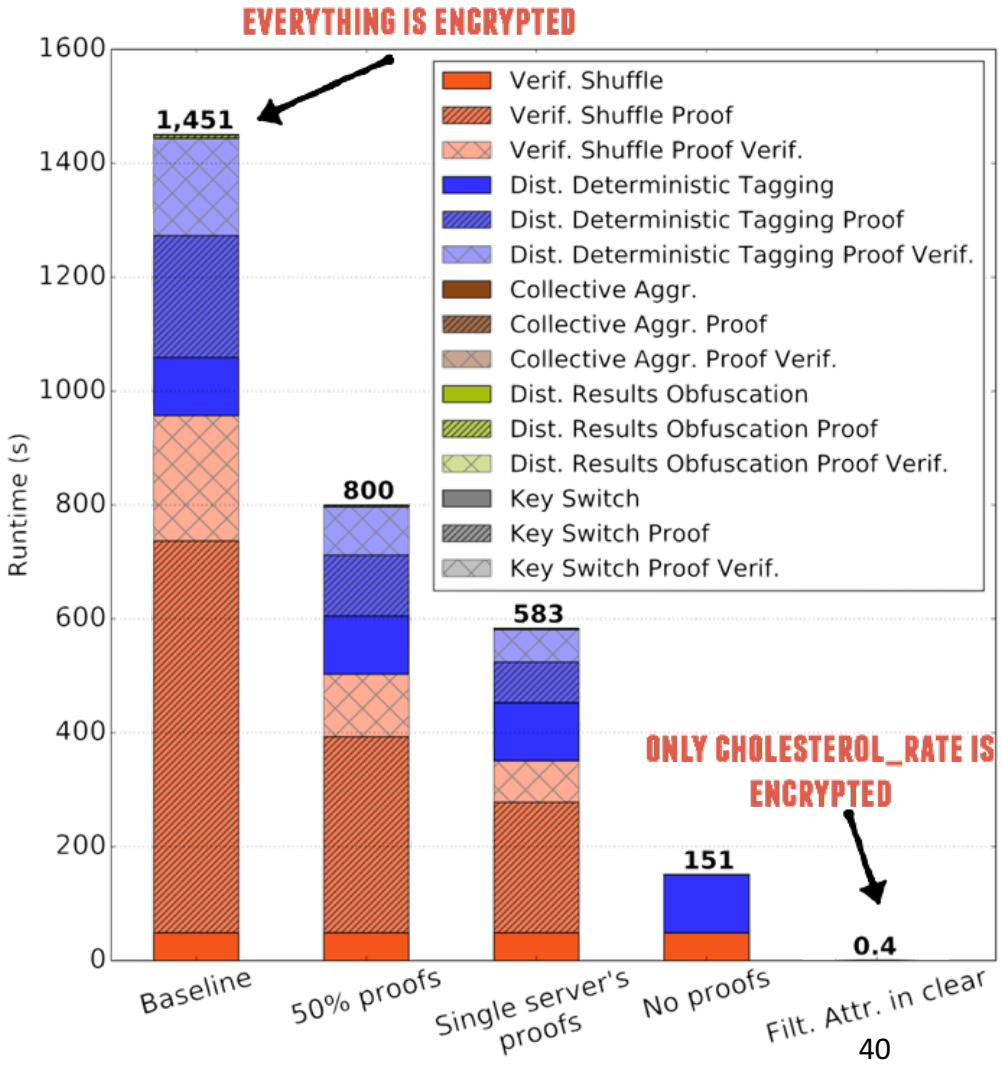
– 128 bit security (using Ed25519 Elliptic Curve)

Performance Evaluation

QUERY

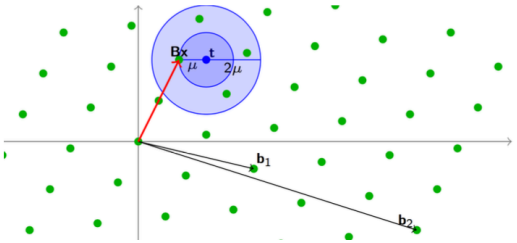
```
SELECT AVG (cholesterol_rate)
FROM DP1,...,DP20
WHERE age in [40:50]
AND ethnicity = Caucasian
GROUP BY gender
```



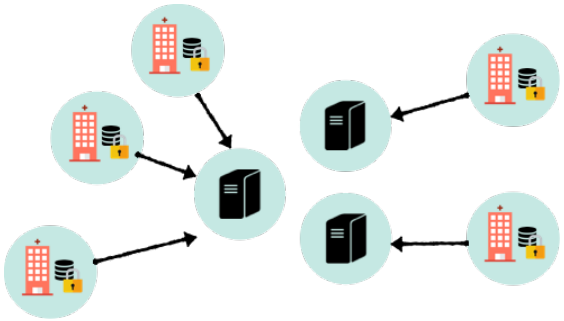


UnLynx Future Developments

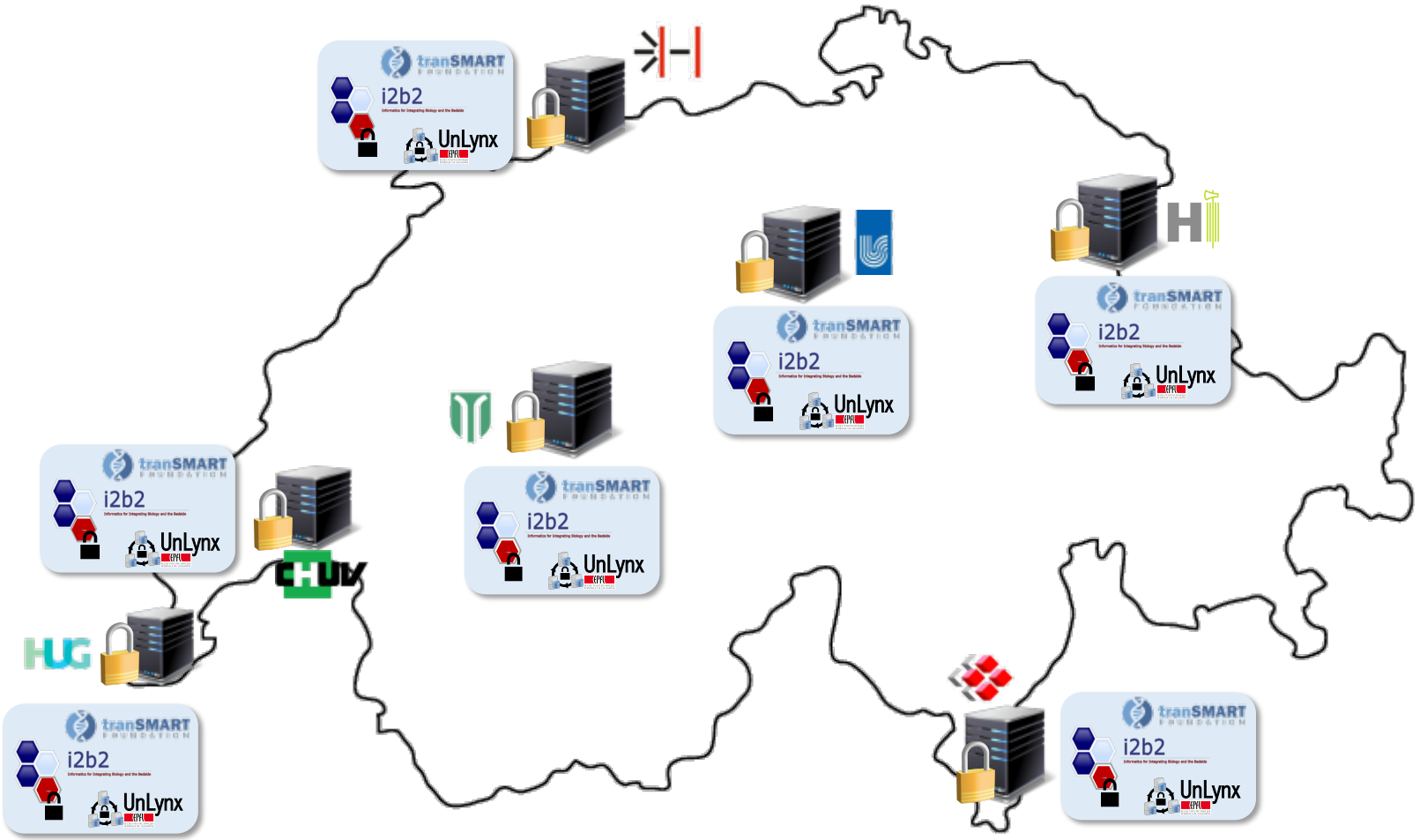
- Change the underlying homomorphic encryption scheme from ElGamal to Lattices to enable multiplications and more complex queries



- Provide accountability guarantees, identity management and topology management through the use of blockchains
- Real-world deployment in a medical use case



Envisioned Nation-Wide Deployment



… with possible international extensions

Fitness-Tracking by Health Insurers (mHealth Sensors and Apps)

NEWS > BUSINESS > HEALTH CARE

Companies making fitness-tracking deals with workers for cheaper insurance

BLOOMBERG

08/21/2014 1:24 PM | Updated: 08/21/2014 1:24 PM

Story Comments

To fight rising medical costs, oil company BP last year offered Cory Slagle – a 260-pound former football lineman – an unusual way to trim $1,200 from his annual insurance bill.

One option was to wear a fitness-tracking bracelet from Fitbit Inc. to earn points toward cheaper health insurance.

Projet pilote myStep: la CSS a une longueur d'avance

Communiqué de presse, le 9 juin 2015

CSS Assurance

La CSS Assurance fait un pas de plus vers la numérisation dans le monde de la santé. En collaboration avec l'Université de Saint-Gall (HSG) et l'EPF de Zurich, elle lance le projet pilote myStep. Pour cette étude scientifique, elle utilise des podomètres afin de déterminer comment il est possible de concevoir de manière optimale une offre de prévention numérique.

Our Main International Research Partners on Protection of -omics Data

- Cornell Tech
- Global Alliance for Genomics and Health (GA4GH)
- Harvard U.
- Longevity, Inc.
- Stanford U.
- UC San Diego
- U. College London
- U. of Darmstadt
- U. of Illinois at Bloomington
- Vanderbilt U.

Events on Genome Privacy and Security

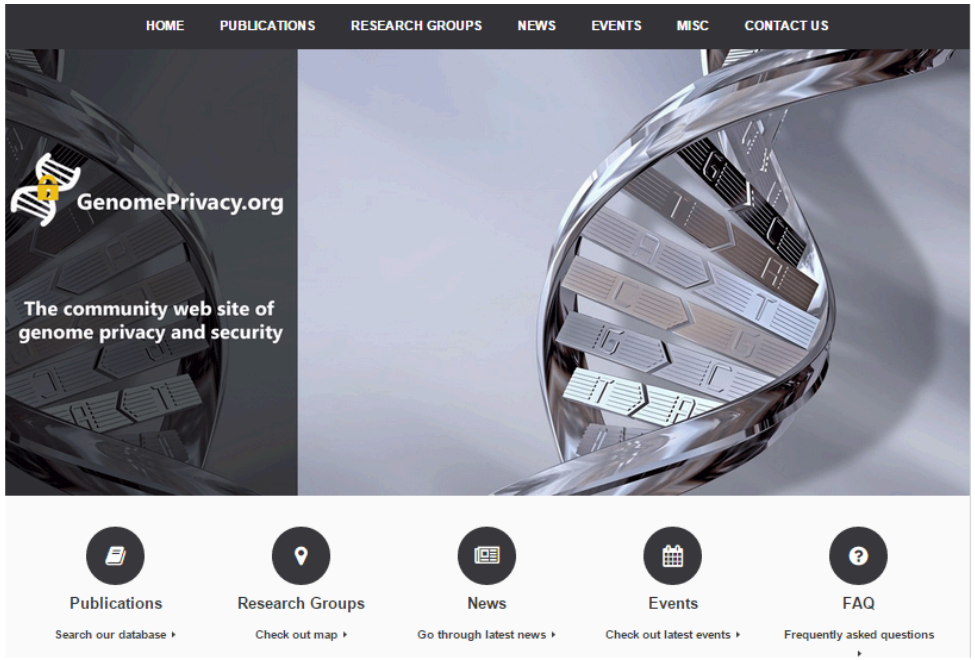
- **Dagstuhl** seminars on genome privacy and security 2013, 2015
- **Conference on Genome and Patient Privacy (GaPP)**
  - **March 2016, Stanford School of Medicine**
- **GenoPri**: International Workshop on Genome Privacy and Security
  - July 2014: Amsterdam (co-located with PETS)
  - May 2015: San 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**

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

iDASH

➔ Lots of material online

“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**
- Precision medicine requires collecting and sharing many more data
- Presence of **genomic data** 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 decentralised tools** (UnLynx)
- We have **operational prototypes**, currently in deployment phase (at Lausanne University Hospital)
- There is a tremendous need for **standardization**, especially for multi-site studies
- Our contributions to genome privacy and security:
  http://lca.epfl.ch/projects/genomic-privacy/