

Privacy & Cyberbiosecurity

Eric Ni

CBB752b22

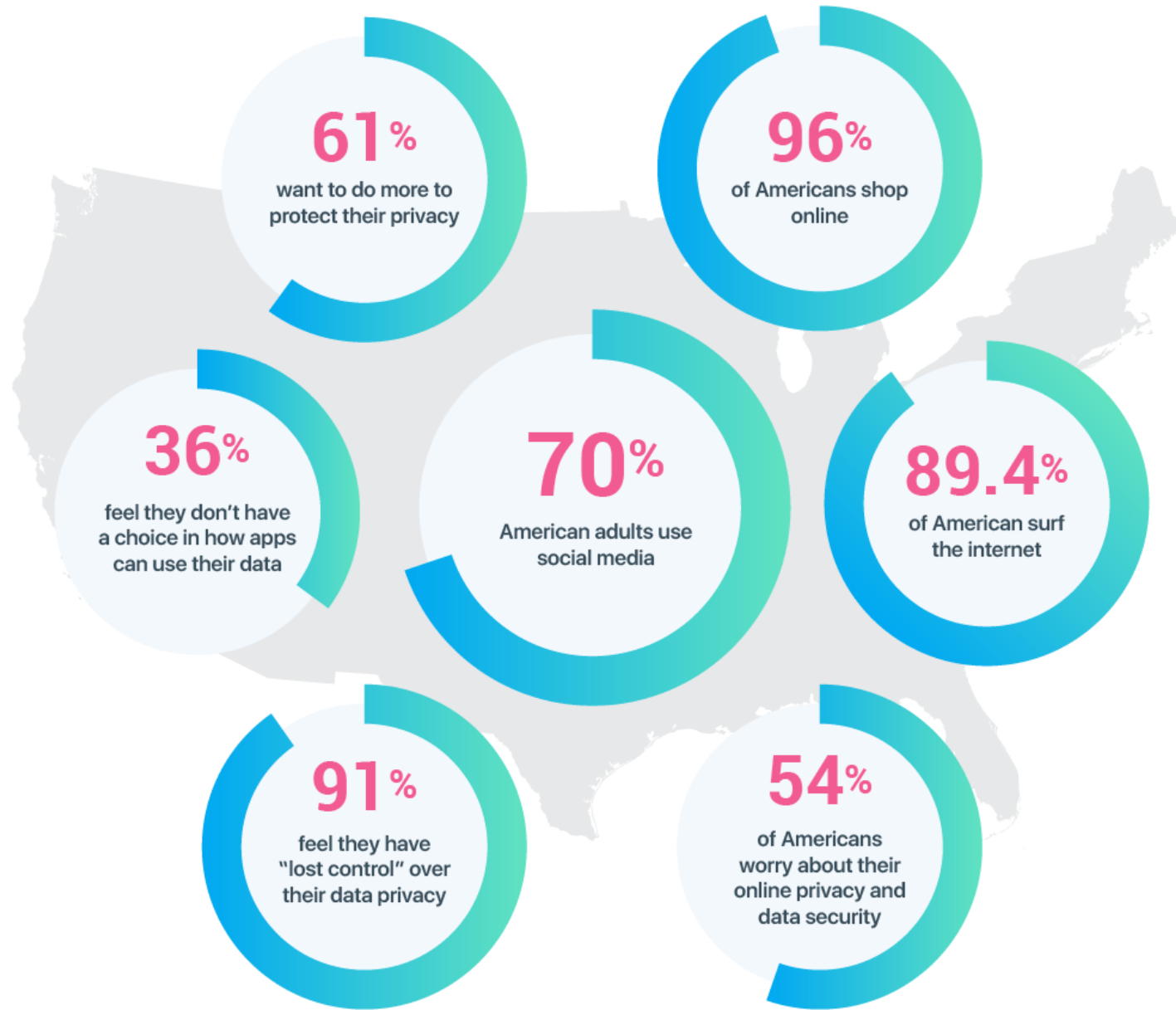
3/14/2022

Outline

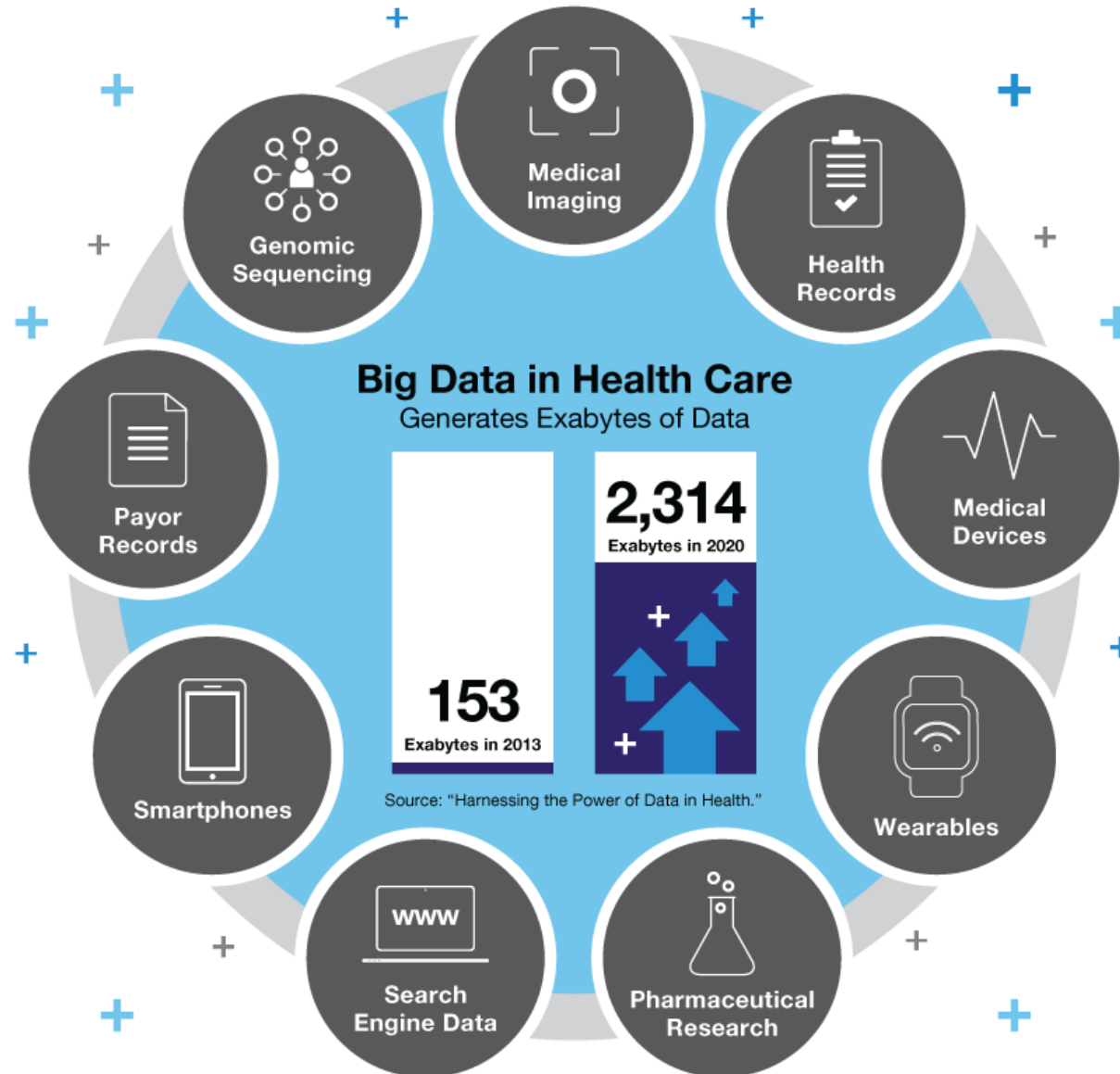
- Privacy & Security
- Cyberbiosecurity
- The Biomedical Data Life Cycle: Vulnerabilities and Countermeasures
 - Data collection, creation, and storage
 - Data analysis and tool development
 - Data dissemination

- See also: dov's lecture from last year:
<http://cbb752b21.gersteinlab.org/syllabus>





Where is all the health care data coming from?



April 26, 2018

U.S.

How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect

Investigators used DNA from crime scenes and plugged that genetic profile into an online genealogy database, tracing DNA to the suspect, Joseph James DeAngelo.

By Thomas Fuller



PRINT EDITION Genealogy Site Led to the Suspect's Front Door | April 27, 2018, Page A19

April 27, 2018

HEALTH

The Golden State Killer Is Tracked Through a Thicket of DNA, and Experts Shudder

The arrest of a suspect has set off alarms among some scientists and ethicists worried that consumer DNA may be widely accessed by law enforcement.

By Gina Kolata and Heather Murphy

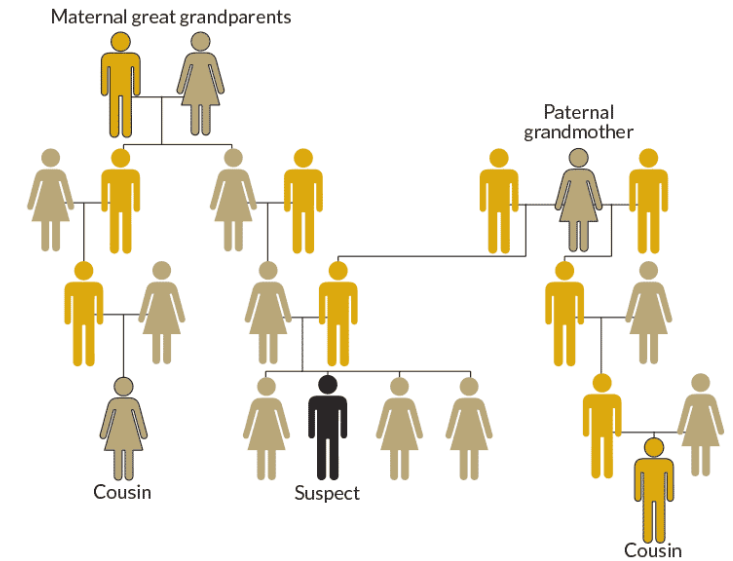


PRINT EDITION Stores of DNA That Anybody Can Pore Over | April 28, 2018, Page A1

Table 2

List of popular DTC companies (in alphabetical order) providing health-related services based on genomic data.

DTC Company	Year Founded	Number of Individuals	Main Services
23andMe (https://www.23andme.com)	2006	>10 Millions	Medical, Genealogical, Personal Ancestry
AncestryDNA (https://www.ancestry.com/dna/)	2002	>16 Millions	Genealogical, Personal Ancestry (Autosomal only)
FamilyTreeDNA (https://www.familytreedna.com)	1999	>1.1 Million	Genealogical, Personal Ancestry (Autosomal only)
<u>GEDmatch (https://www.gedmatch.com)</u>	2010	>1.3 Million	Genetic Genealogy Search
MyHeritage (https://www.myheritage.com)	2003	>3 Million	Genealogical, Personal Ancestry (Autosomal only)



Snapshot Prediction Results Phenotype Report



Case #87-9340



Contact: Snohomish County S.O.
Tips: (425) 388-3845

Sex: Male ♂

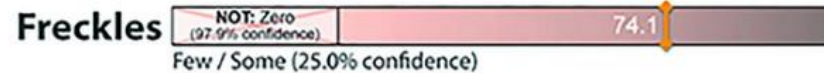
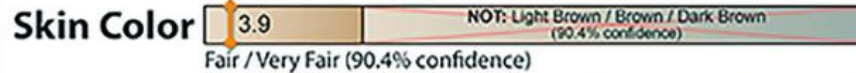
Age: Unknown
(Age progressed to ~65 years)

Body Mass: Unknown
(Shown at BMI 24)

Ancestry: Northern European



Region	Percent
Europe - North	90.81%






© 2018 Parabon NanoLabs, Inc. All rights reserved.

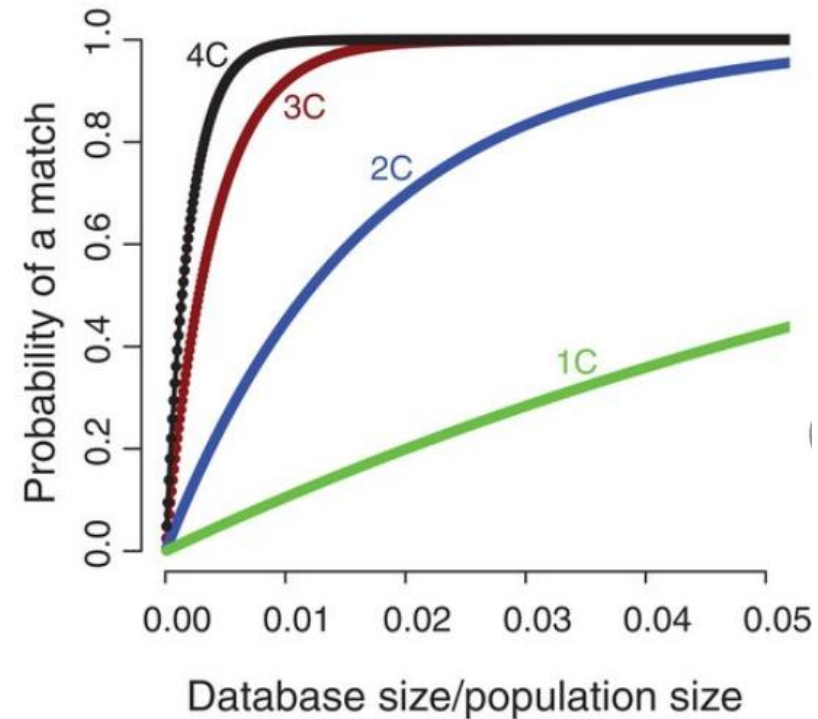
<https://Parabon-NanoLabs.com/Snapshot>

Parabon, a forensic consulting firm, generated this composite sketch of the suspect using crime-scene DNA.
Snohomish County Sheriff's Office, via Associated Press

Identity inference of genomic data using long-range familial searches

YANIV ERLICH  , TAL SHOR  , ITSIK PE'ER  , AND SHAI CARMİ 

SCIENCE • 11 Oct 2018 • Vol 362, Issue 6415 • pp. 690-694 • DOI: [10.1126/science.aau4832](https://doi.org/10.1126/science.aau4832)



Privacy defined

- “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others” – Alan Westin (1967)
- Ownership: “‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’)” – GDPR (2018)

Privacy is different than Security

Privacy:

Control over personal information

The right to be let alone

Limited access to the self

Secrecy

Personhood/Ownership

Intimacy

Security:

Confidentiality

Integrity

Availability

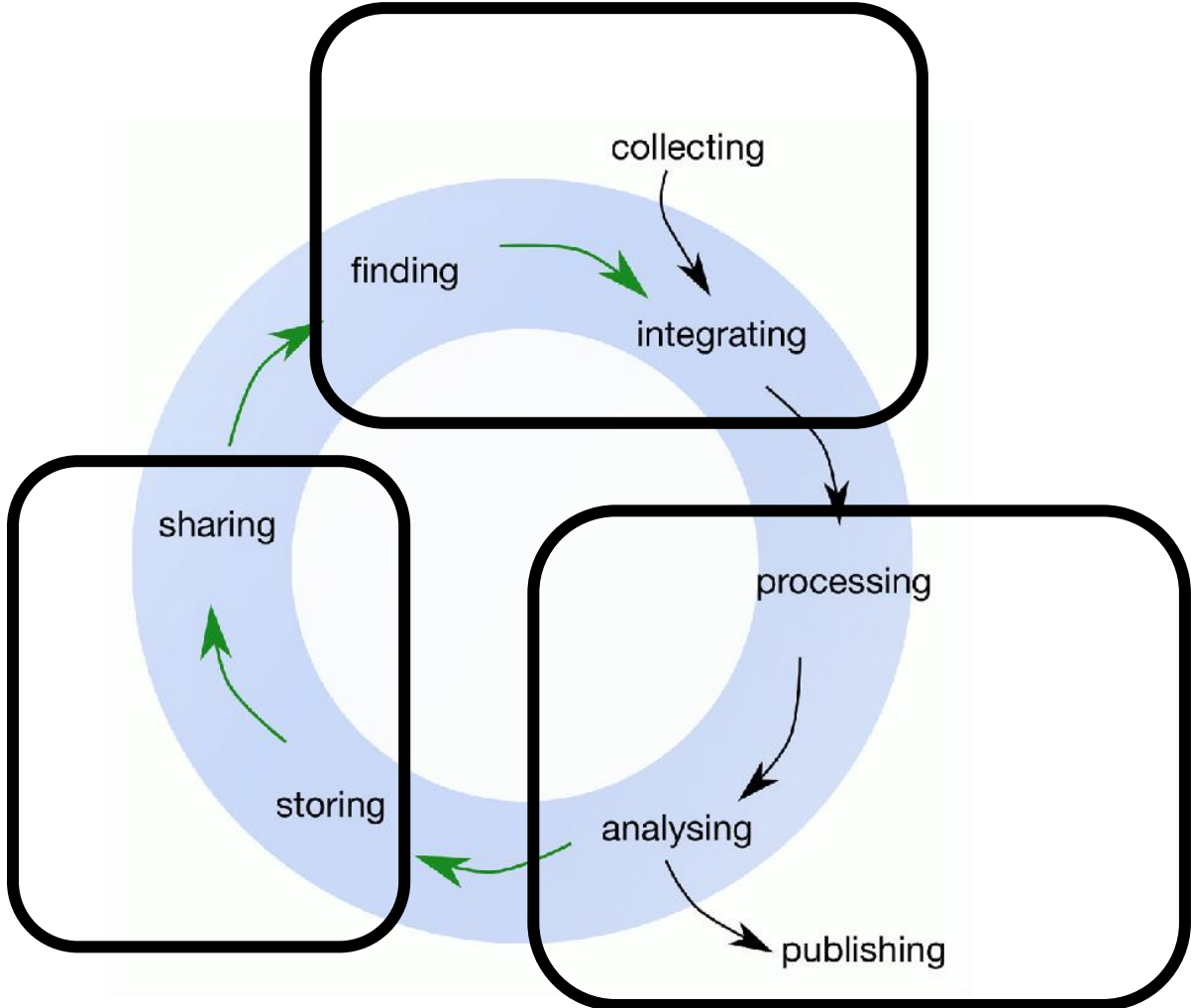
Reliability

Authenticity

Cyberbiosecurity

- **Biosecurity:** securing valuable biological material from misuse or harm
- **Cybersecurity:** protection of computer systems from theft and damage to their hardware, software, or information, as well as from disruption or misdirection of the services they provide
- **Cyberbiosecurity:** addresses the potential for or actual malicious destruction, misuse, or exploitation of valuable information, processes, and material at the interface of the life sciences and digital worlds

The Biomedical Data Lifecycle



Data collection, creation, and storage

Vulnerabilities

- Data theft/unauthorized access
 - Phishing
 - Malware/ransomware
 - Social engineering
- Data integrity
 - Loss of data
 - Manipulation of data
- Central point of failure
- Ownership

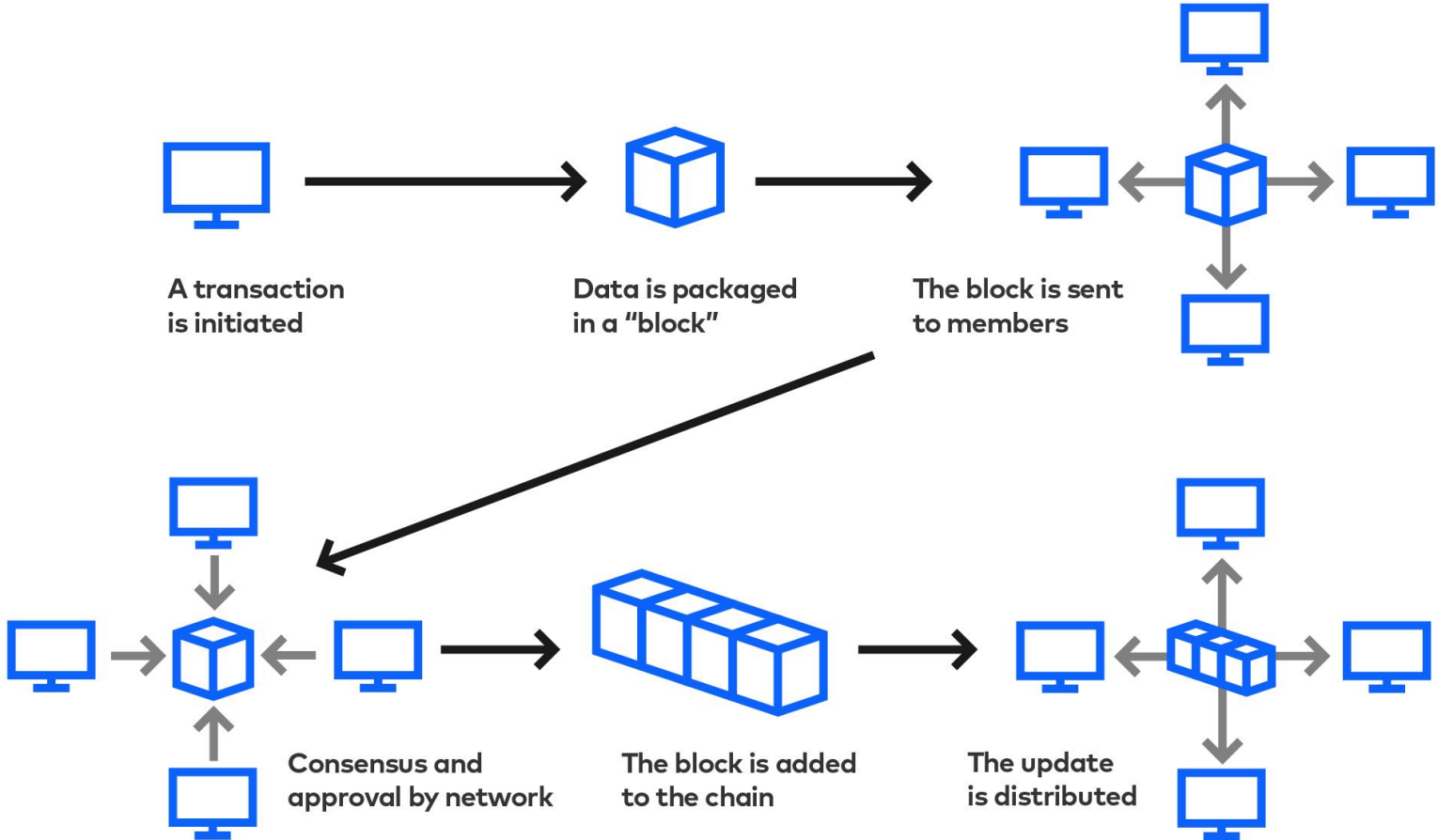
Solutions

- Cybersecurity protocols
- Law/policy
- Establish standards & frameworks
- Decentralized storage
 - Blockchain
 - IPFS

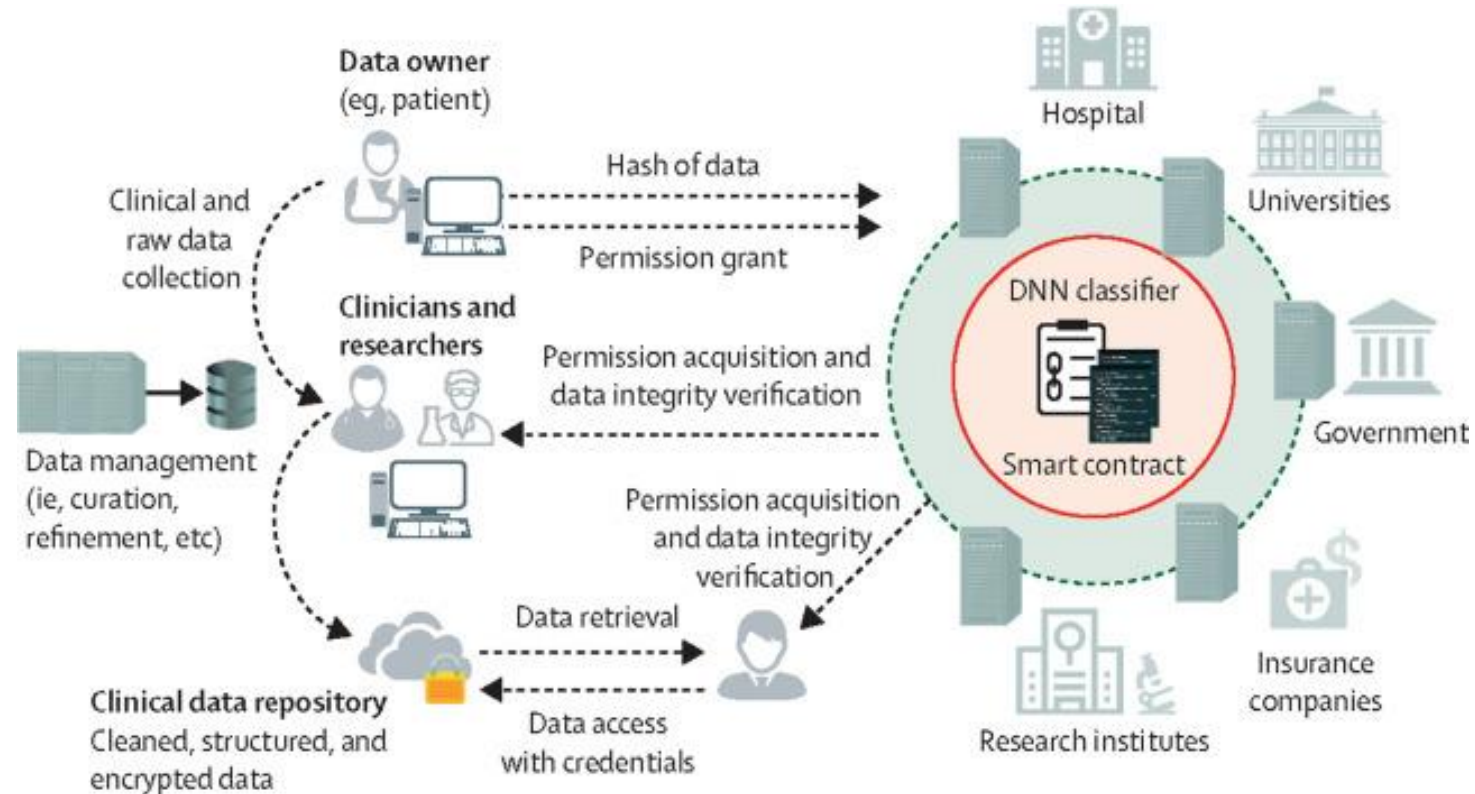
Who owns your health data?

- Legally, varies by state, but usually, not the patients
 - In most states, legal ownership still resides in your healthcare provider
 - Hard to define 'ownership' generally
- HIPAA protects this data, and patients can “inspect, review and receive a copy of his or her own medical records and billing records”

Blockchain



Blockchain: a solution for EHR sharing



Data analysis & tool development

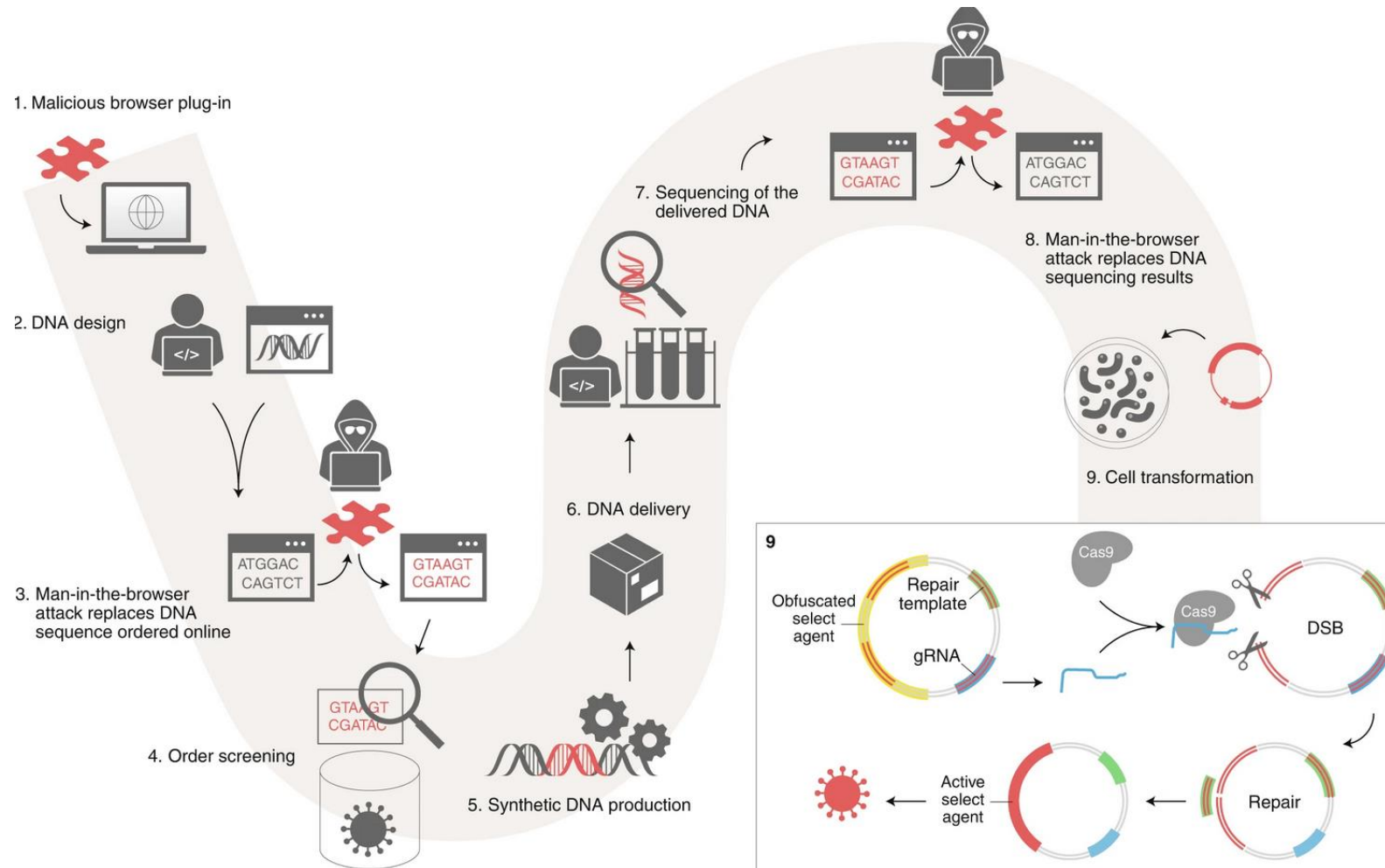
Vulnerabilities

- Improper use of model
 - Adversarial attacks
 - Data poisoning
 - DNA injection attack
- Privacy leakage from model
 - Model inversion attack

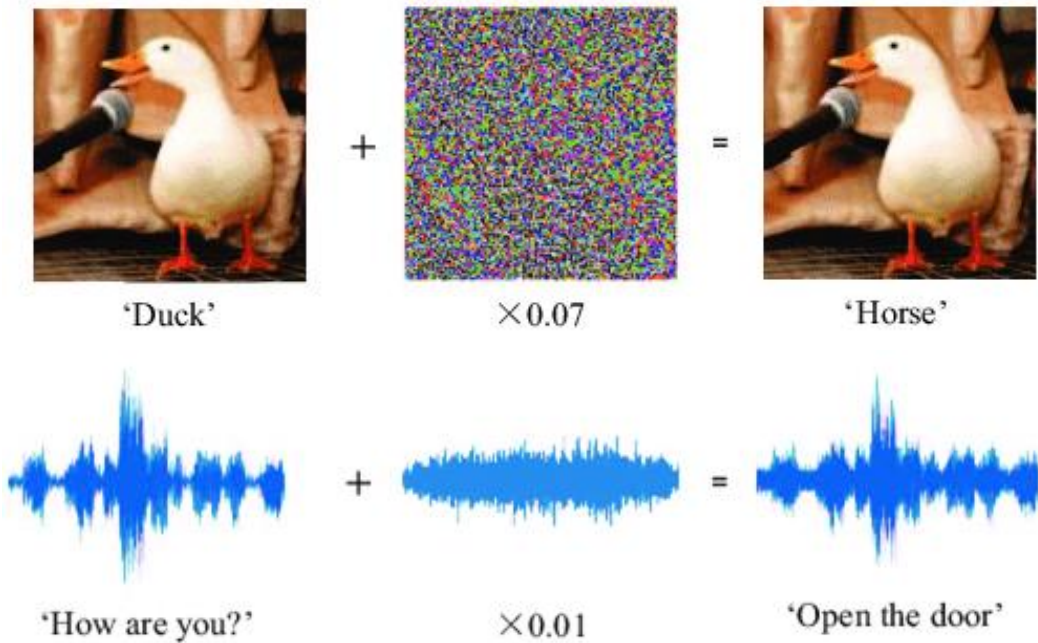
Solutions

- Generalization
- Regularization
- Differential Privacy

Malicious DNA injection attack



Adversarial attacks



Adversarial Examples



Clean Stop Sign

"Stop sign"



Real-world Stop Sign
in Berkeley

"Stop sign"



Adversarial Example

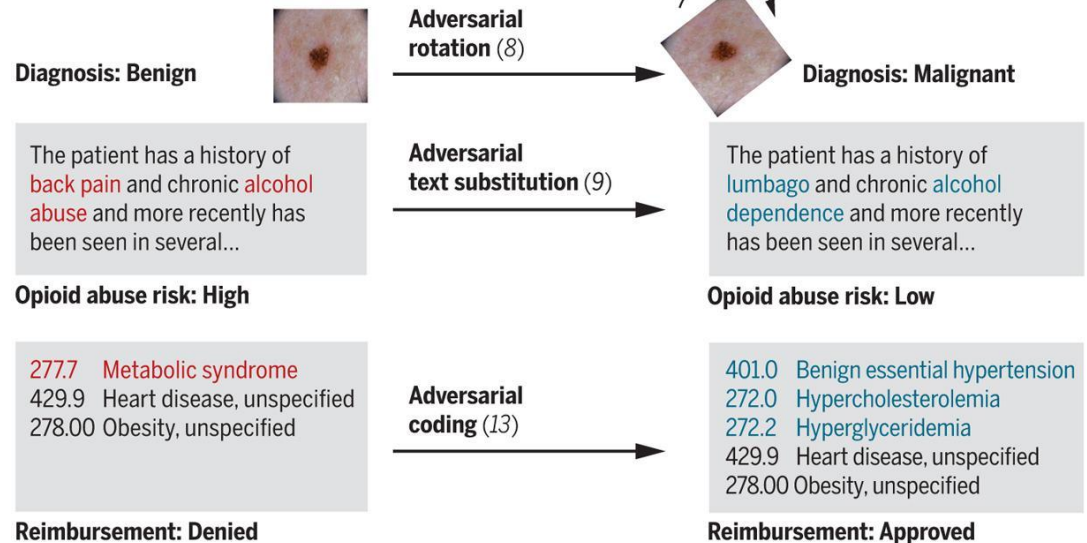
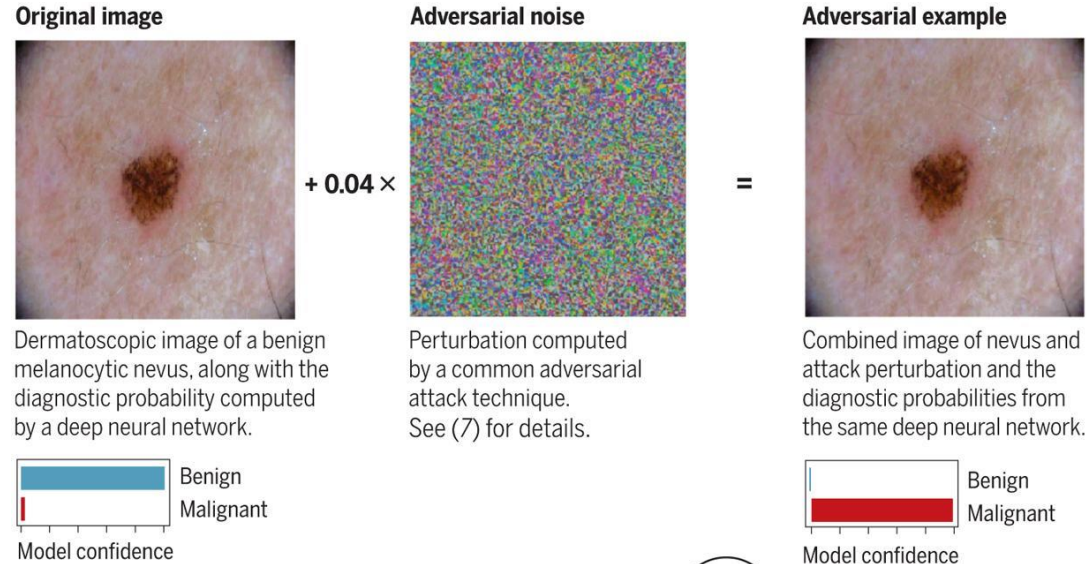
"Speed limit sign
45km/h"



Adversarial Example

"Speed limit sign 45km/h"

Adversarial attacks



Data poisoning in SVM

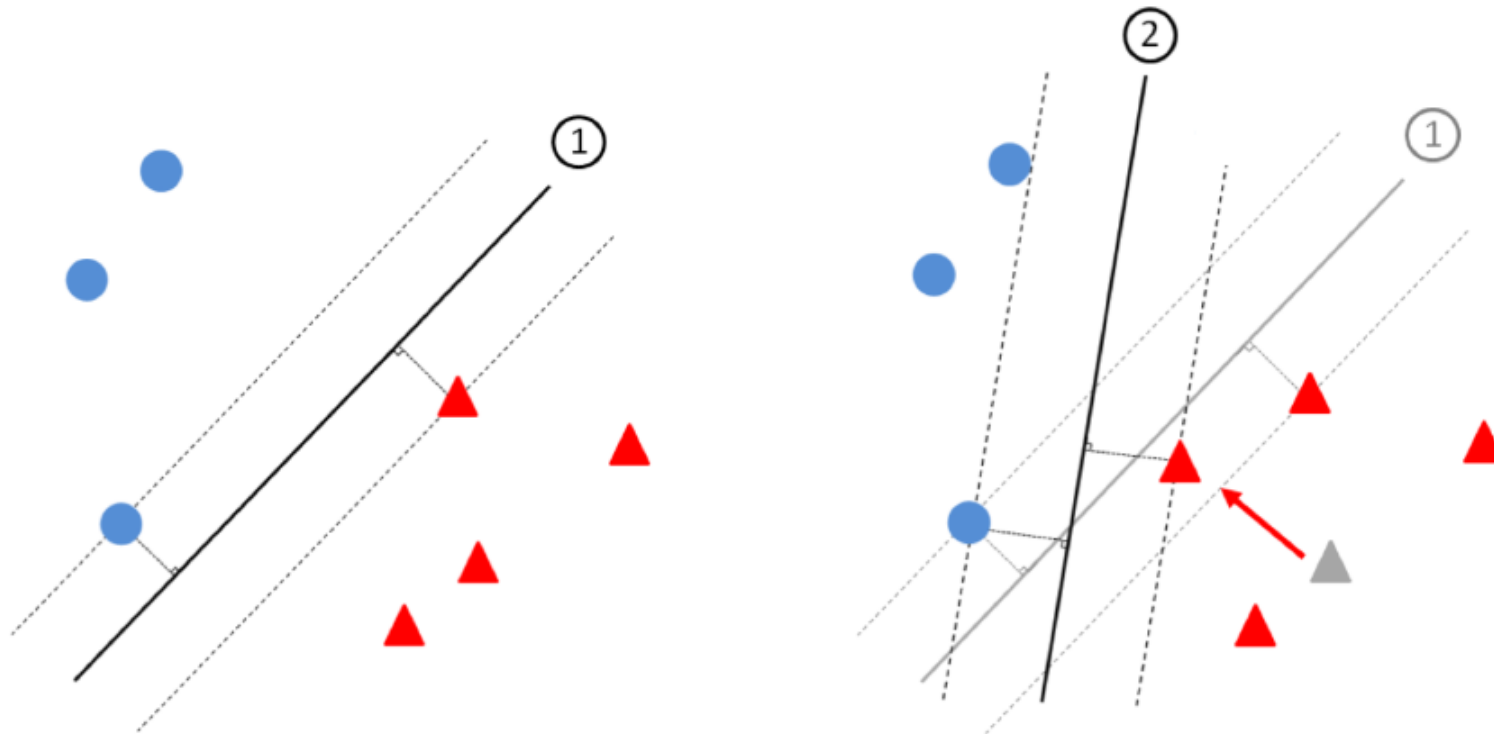


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted in this example if just one training sample is changed, even when that sample's class label does not change (right).

Model inversion attacks recover training data

age	height	weight	race	history	vkorc1	cyp2c9	dose
50-60	176.2	185.7	asian	cancer	A/G	*1/*3	42.0

adversary $\mathcal{A}^f(\text{err}, p_i, x_2, \dots, x_t, y)$:

- 1: for each possible value v of x_1 do
- 2: $x' = (v, x_2, \dots, x_t)$
- 3: $r_v \leftarrow \text{err}(y, f(x')) \cdot \prod_i p_i(x_i)$
- 4: Return $\arg \max_v r_v$

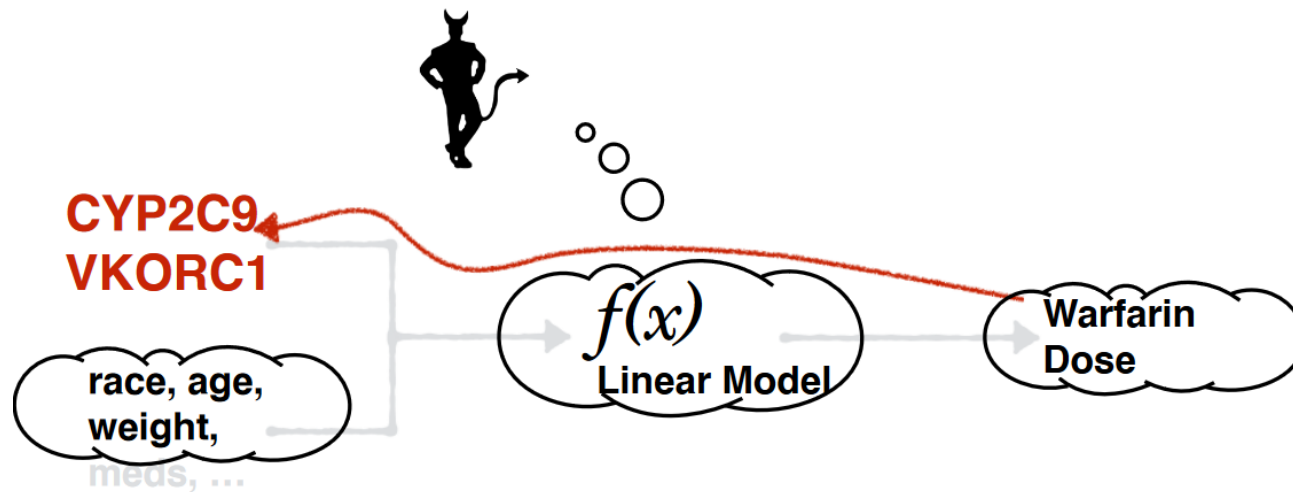
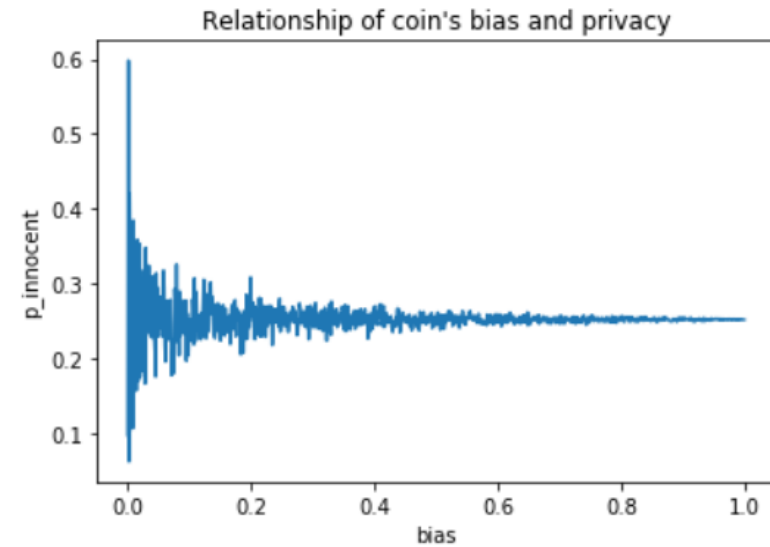
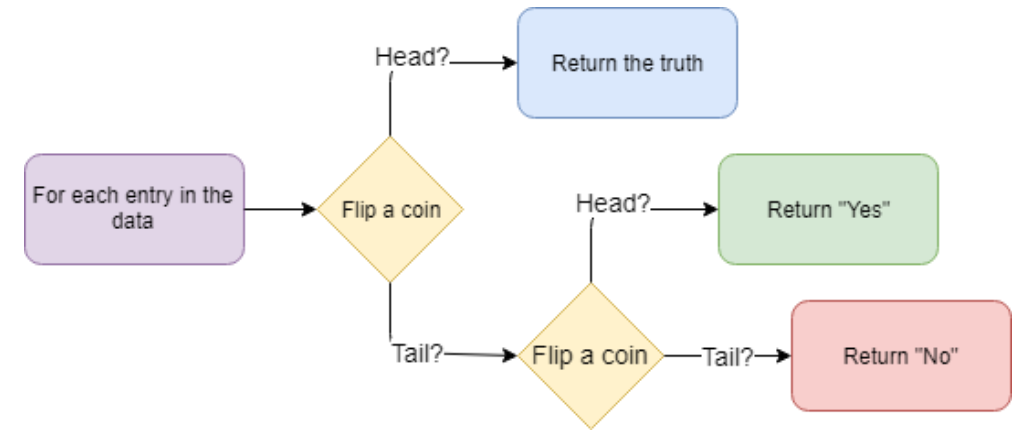
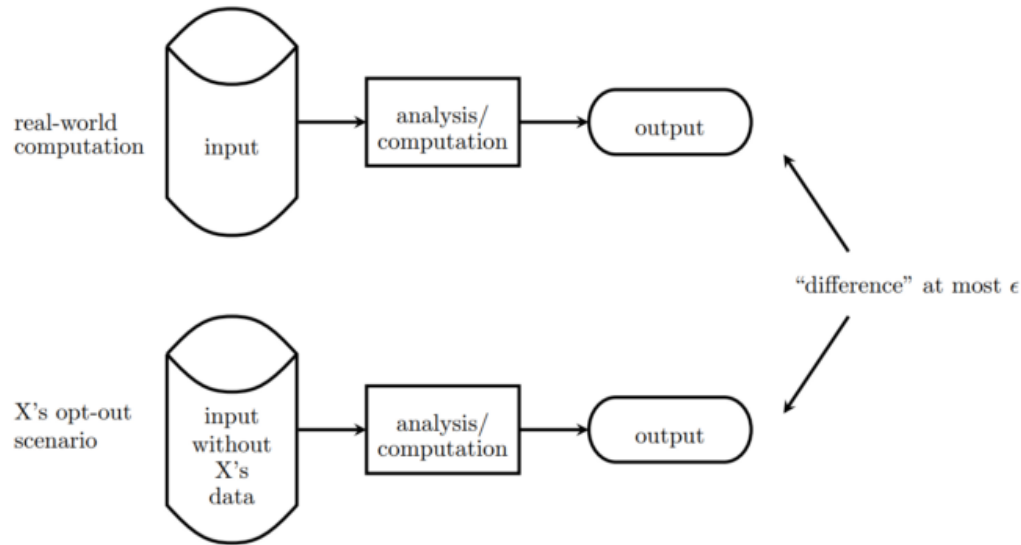


Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Differential privacy

A mathematical definition for privacy that provides a provable guarantee for the degree of privacy protection



Data dissemination

Vulnerabilities

- What data is safe to share?
- Re-identifiability
 - Linking attack
- Privacy leakage

Solutions

- Establish standards & frameworks for sharing data
- Sanitization methods
- Cryptography
 - Secure multiparty computation
 - Homomorphic encryption

HIPAA PHI for de-identification

1. Names;
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death;
4. Phone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social Security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code

Linking Attacks: Case of Netflix Prize



Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...
...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...
...
...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases

Anonymized Netflix Prize Training Dataset
made available to contestants

Linking Attacks: Case of Netflix Prize



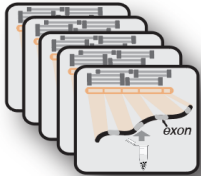
User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
...
...

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
...
...
...

- Many users are shared
- The grades of same users are correlated
- A user grades one movie around the same date in two databases
- IMDB users are public
- NetFLIX and IMdB moves are public

Linking attack: genotype can be linked to reveal phenotypes

Noisy attacked database \mathcal{D} :

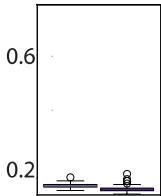


Noisy data as information \mathcal{I} :

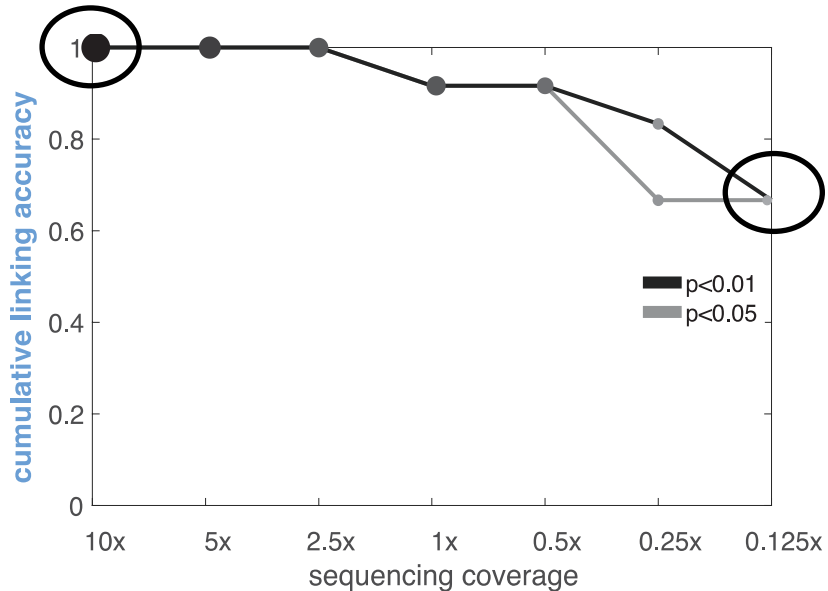
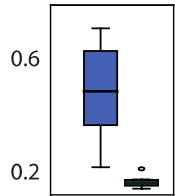
Coffee cups



precision & sensitivity



precision & sensitivity



Privacy leakage in functional genomics

On Sharing Quantitative Trait GWAS Results
in an Era of Multiple-omics Data and the Limits
of Genomic Privacy

Hae Kyung Im,^{1,*} Eric R. Gamazon,² Dan L. Nicolae,^{2,3,4} and Nancy J. Cox^{2,3,*}



The American Journal of Human Genetics 90, 591–598, April 6, 2012

Bayesian method to predict individual SNP genotypes
from gene expression data

Eric E Schadt^{1,5}, Sangsoon Woo^{2,4,5} & Ke Hao^{1,3,5}

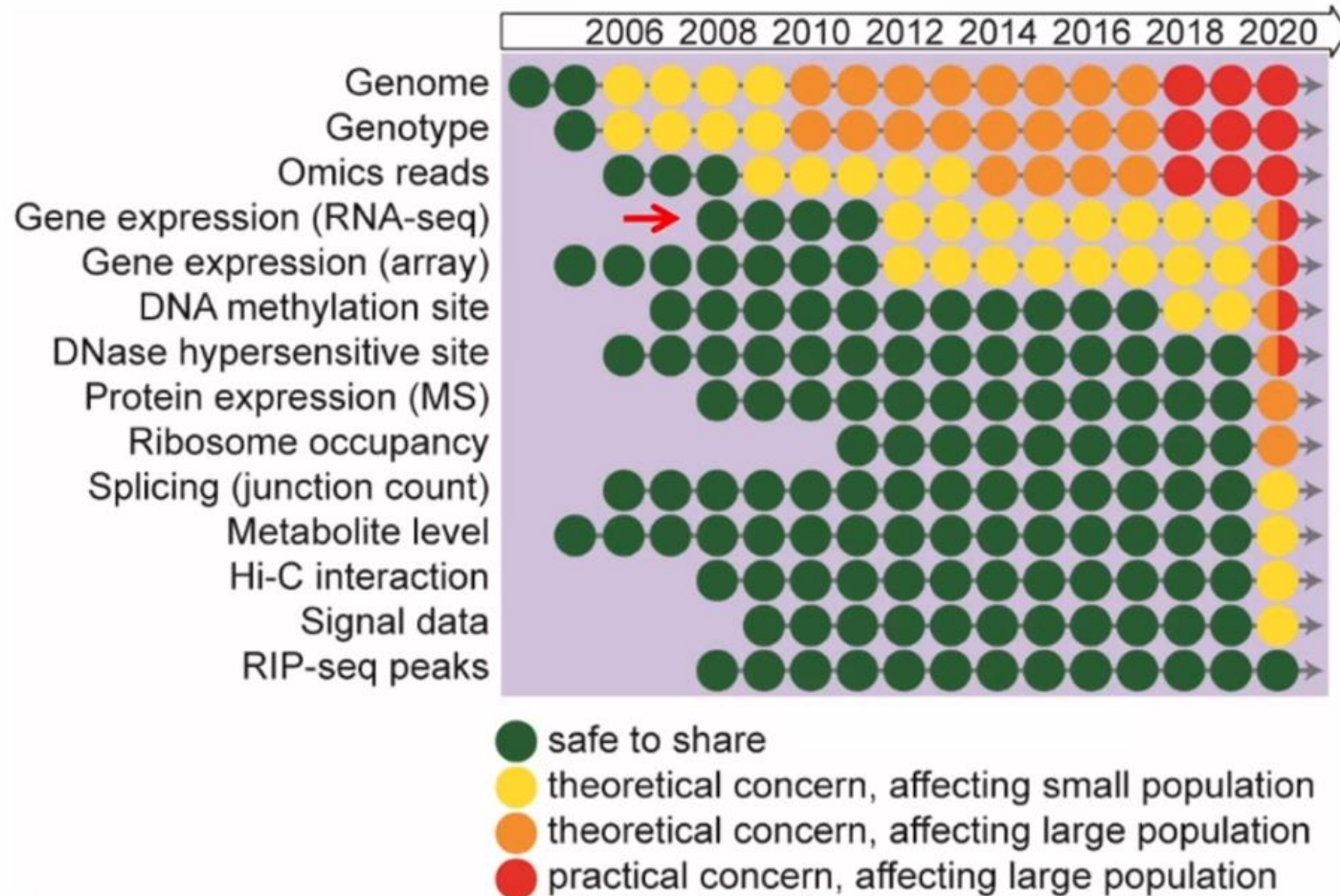
NATURE GENETICS | VOLUME 44 | NUMBER 5 | MAY 2012

Analysis of sensitive information leakage in
functional genomics signal profiles through
genomic deletions

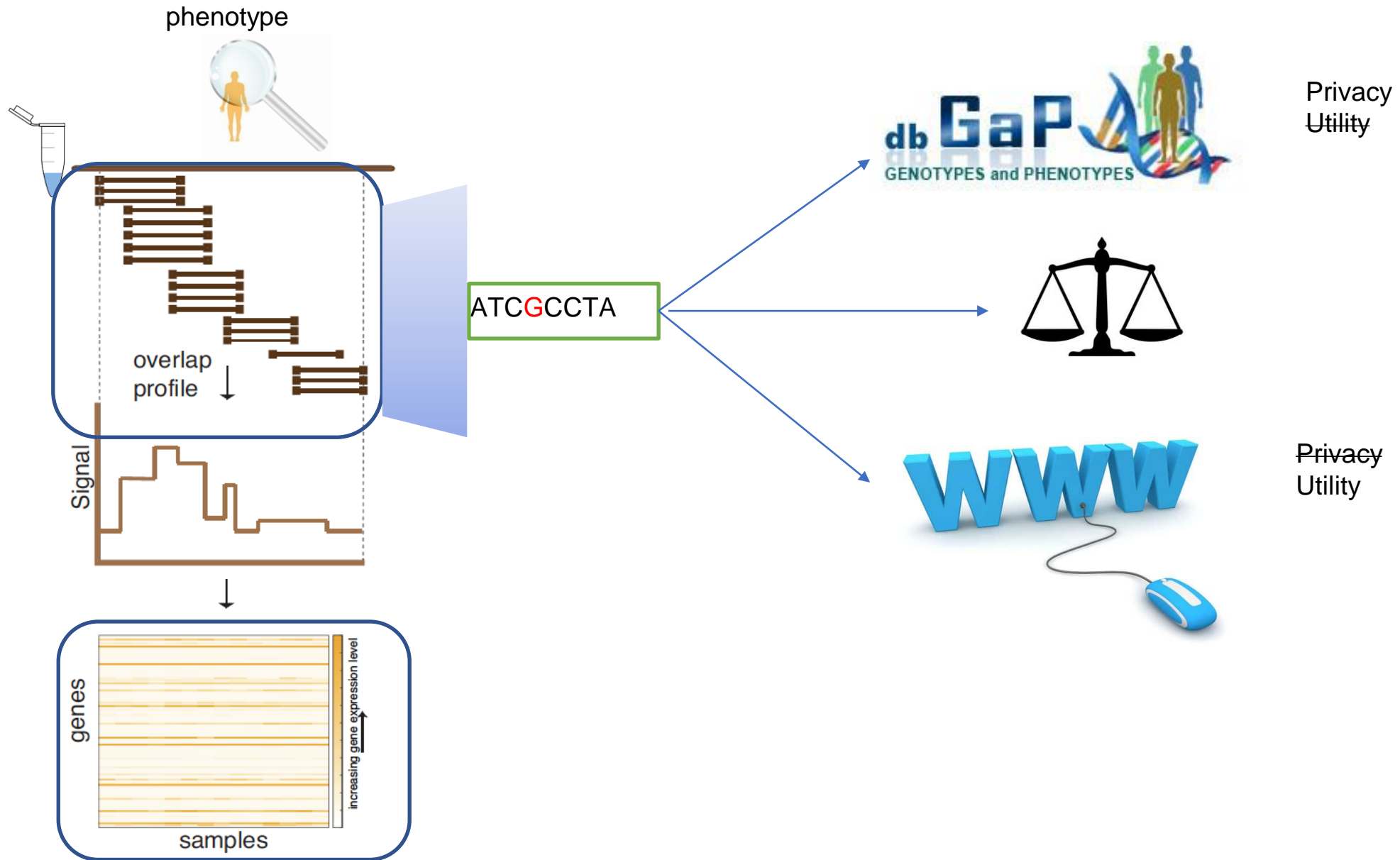
Arif Harmanci ^{1,2,3} & Mark Gerstein ^{1,2,4}

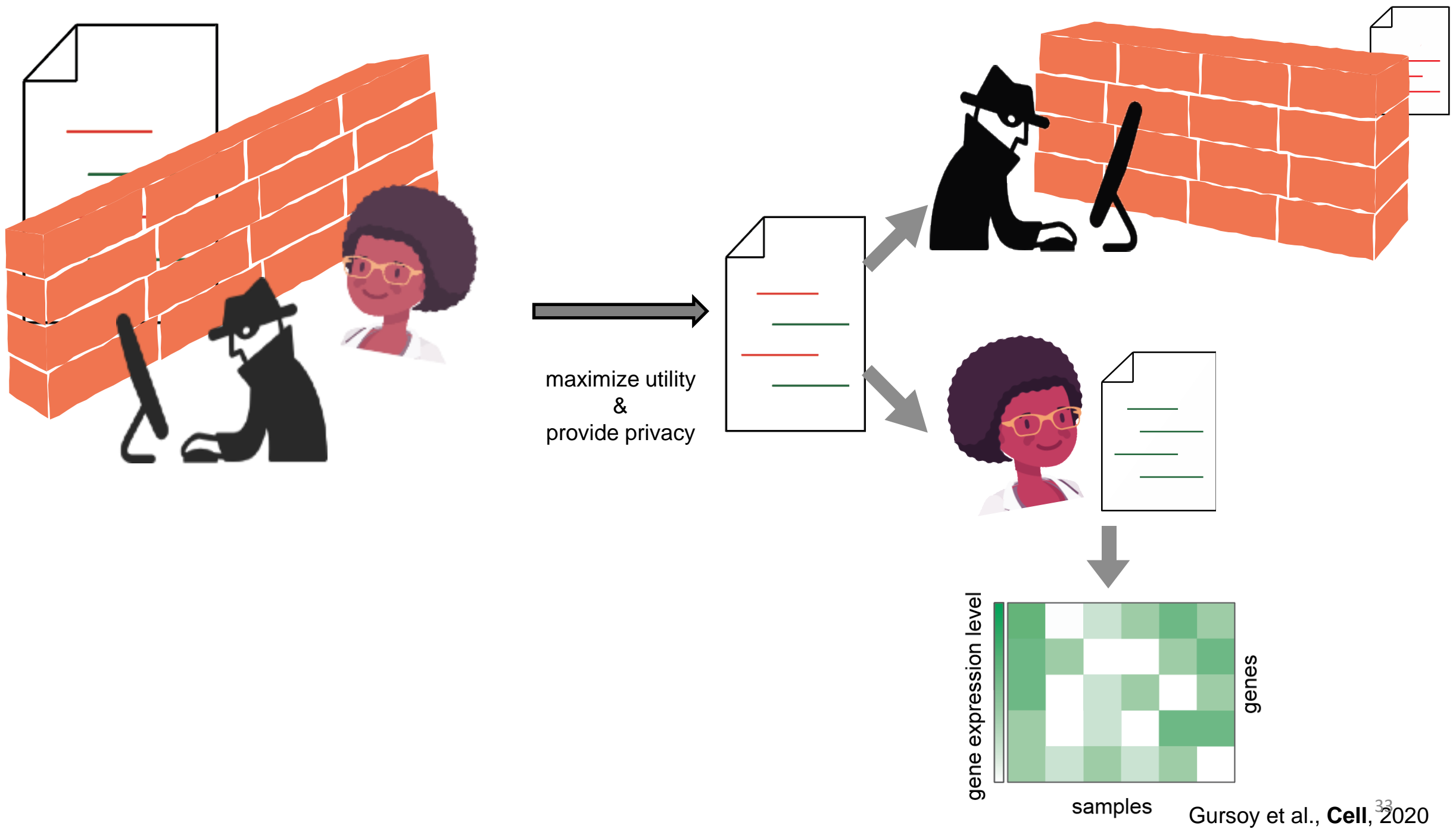
NATURE COMMUNICATIONS | (2018)9:2453 |

Latent functional risk in genomics data manifests over time

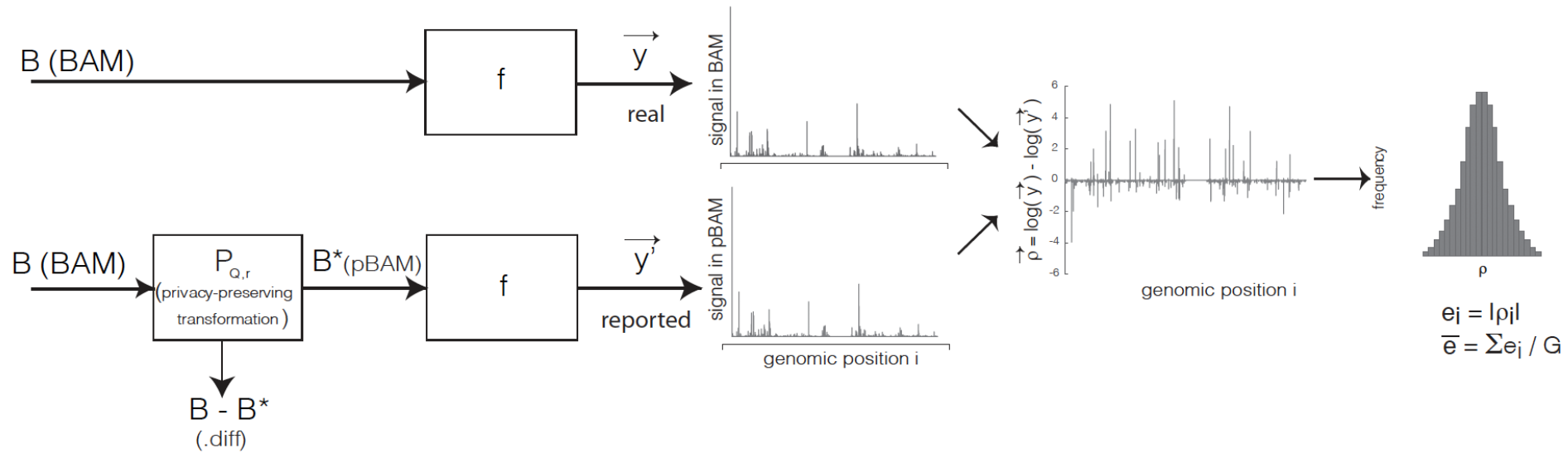


Functional Genome Privacy: a **cautionary** tale?

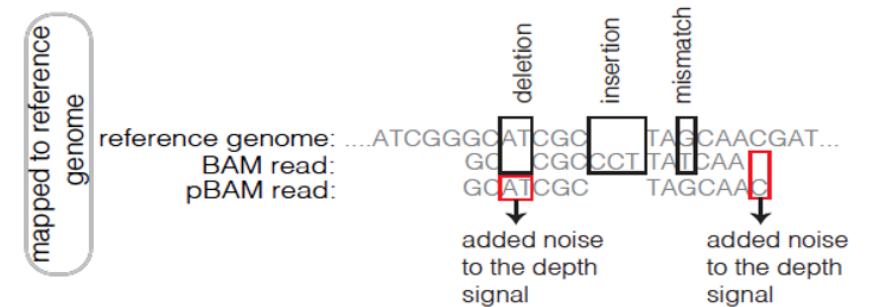




Privacy-preserving Binary Alignment Mapping (pBAM)



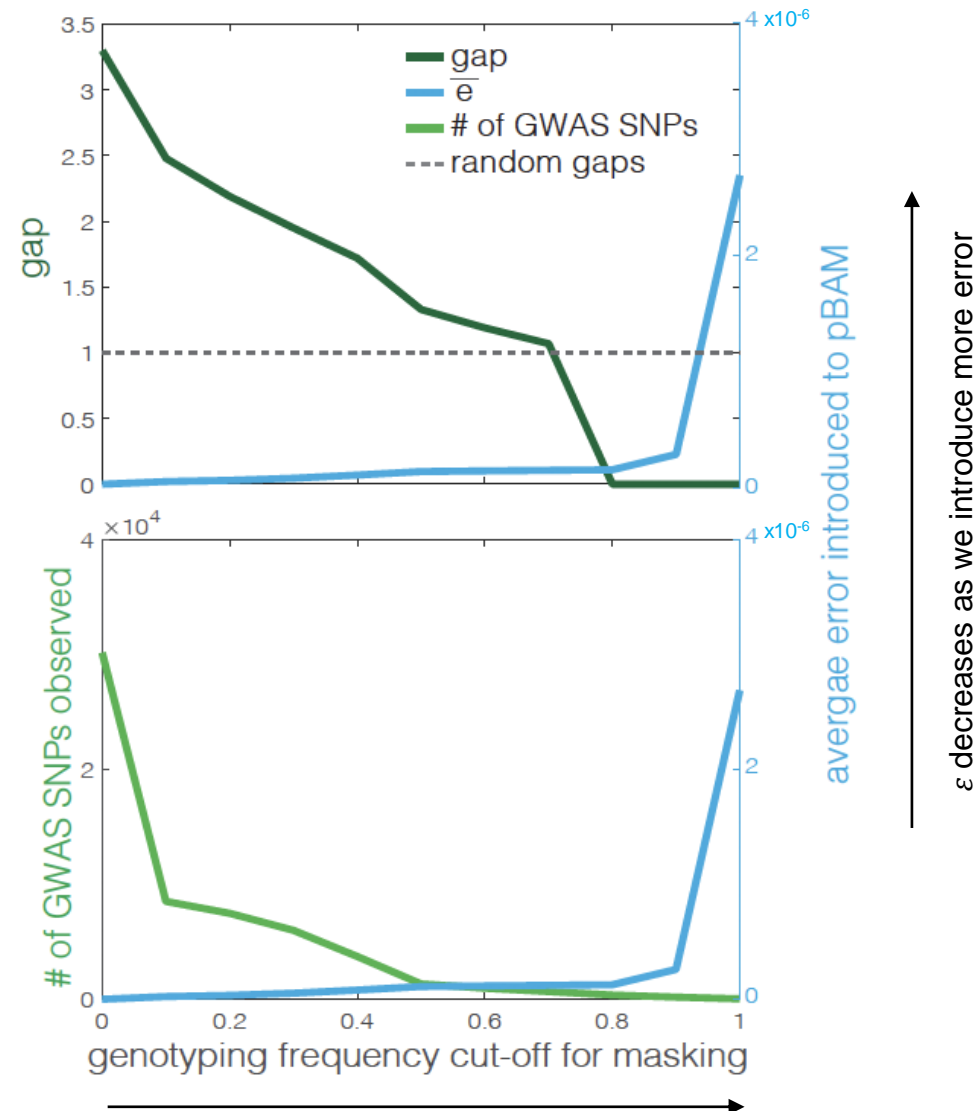
- No need to know the sequence of mapped reads to aggregate them
- A manipulation on Binary Alignment Files (BAM)
 - Find leaky fields/tags
 - Generalization
- Goal:
 - Accurate gene/transcript expression quantification
 - Works with the pipelines / SAMtools



Privacy-preserving Binary Alignment Mapping (pBAM)

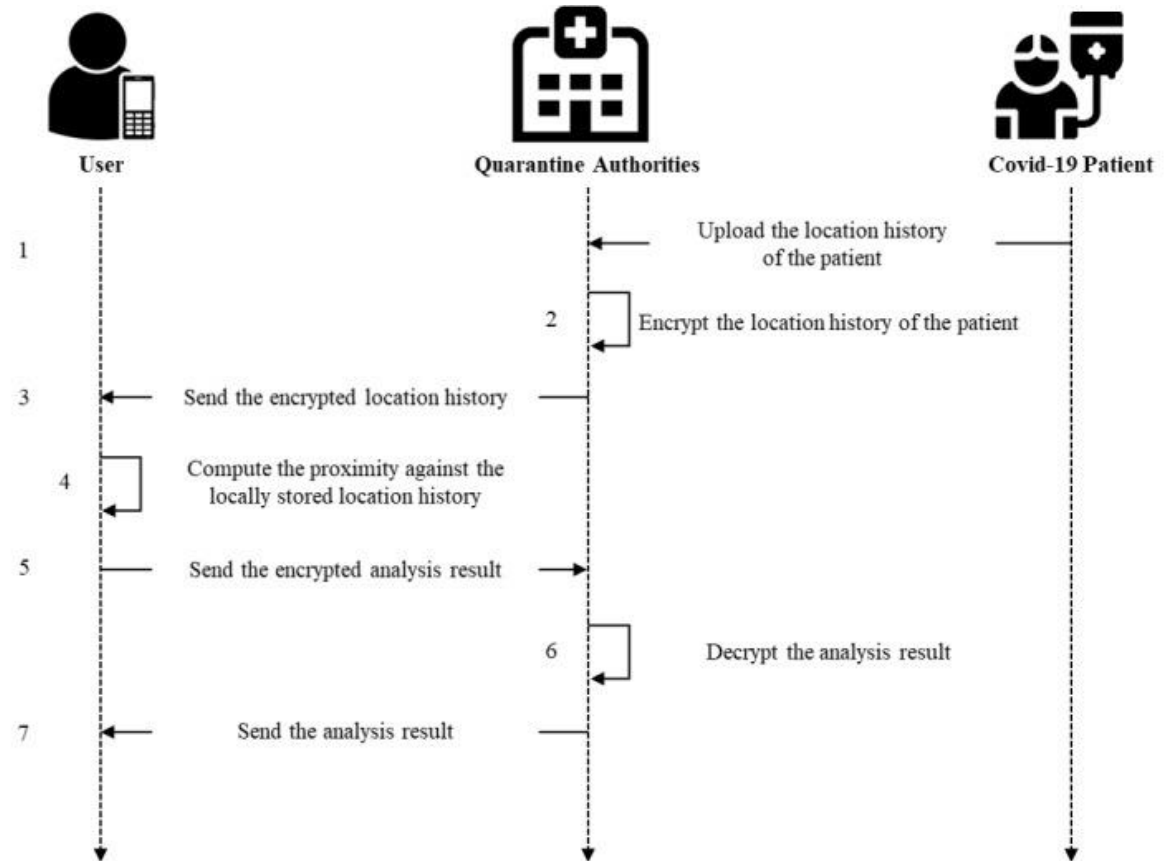
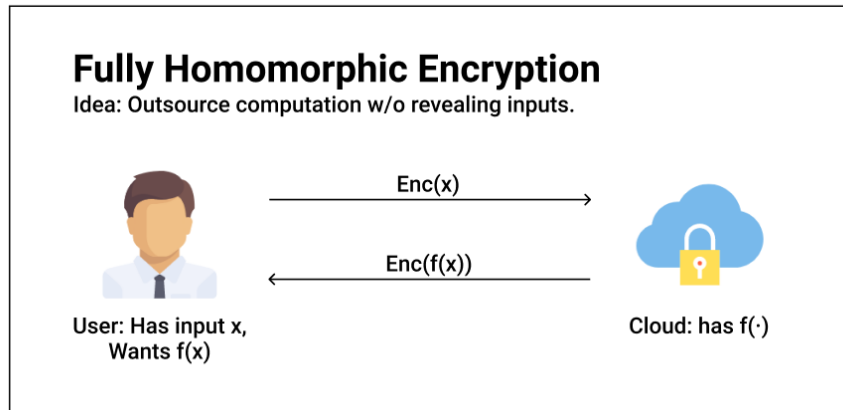
(grounded in privacy and utility)

- Unit = nucleotide (signal track)
- NA12878 RNA-Seq data
- Test the **privacy** for each level of masking
- Measure the **error** introduced



δ increases as we mask more and more common variants

Homomorphic encryption



Summary

- Biomedical Data Science is a fast moving field: with every new analyses and framework come new vulnerabilities
- Policy and laws governing such are lagging far behind scientific discovery
- Sharing data is essential for the progress of medical research, but we cannot share too freely
- Solutions exist to mitigate privacy risk while allowing for research to continue