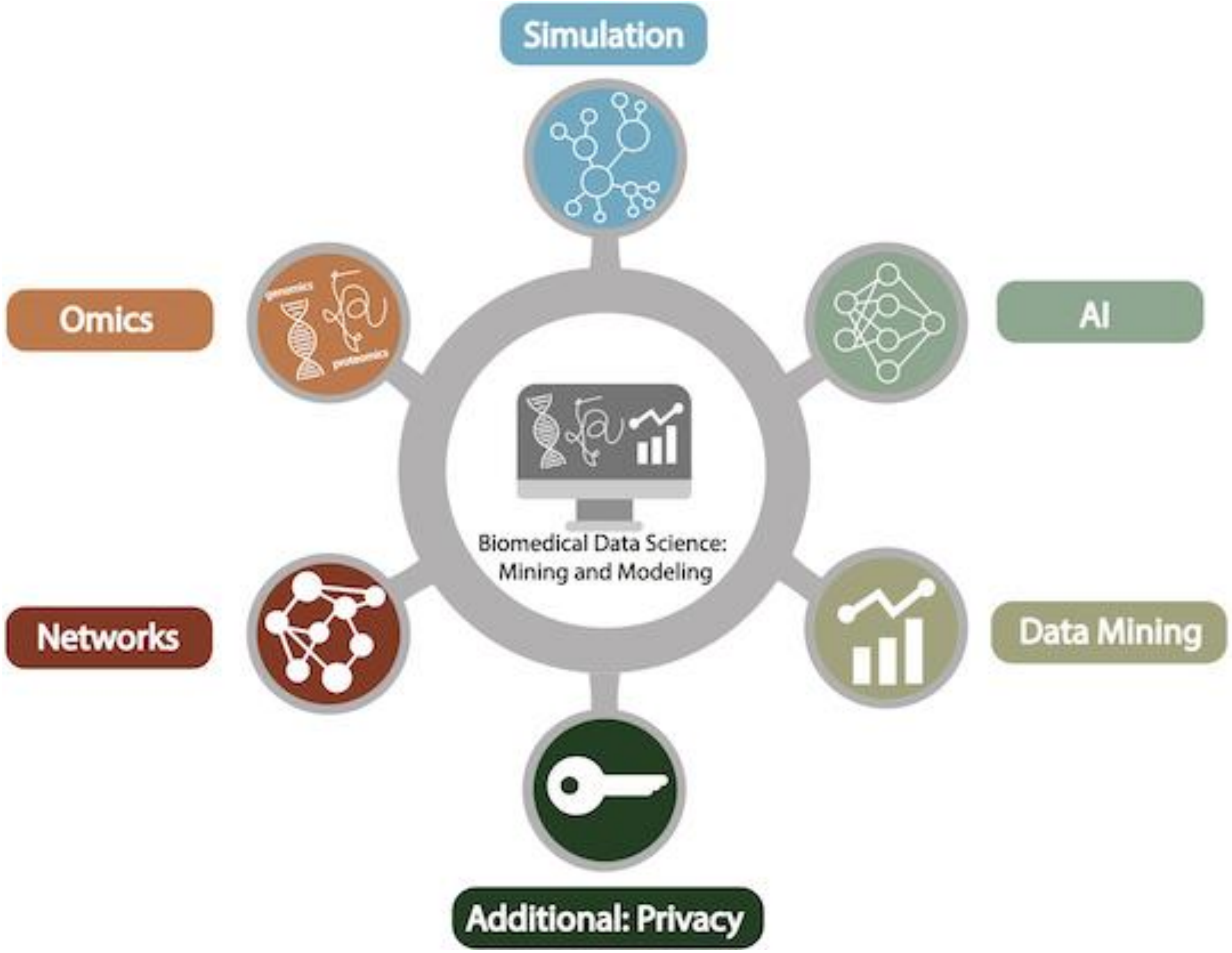
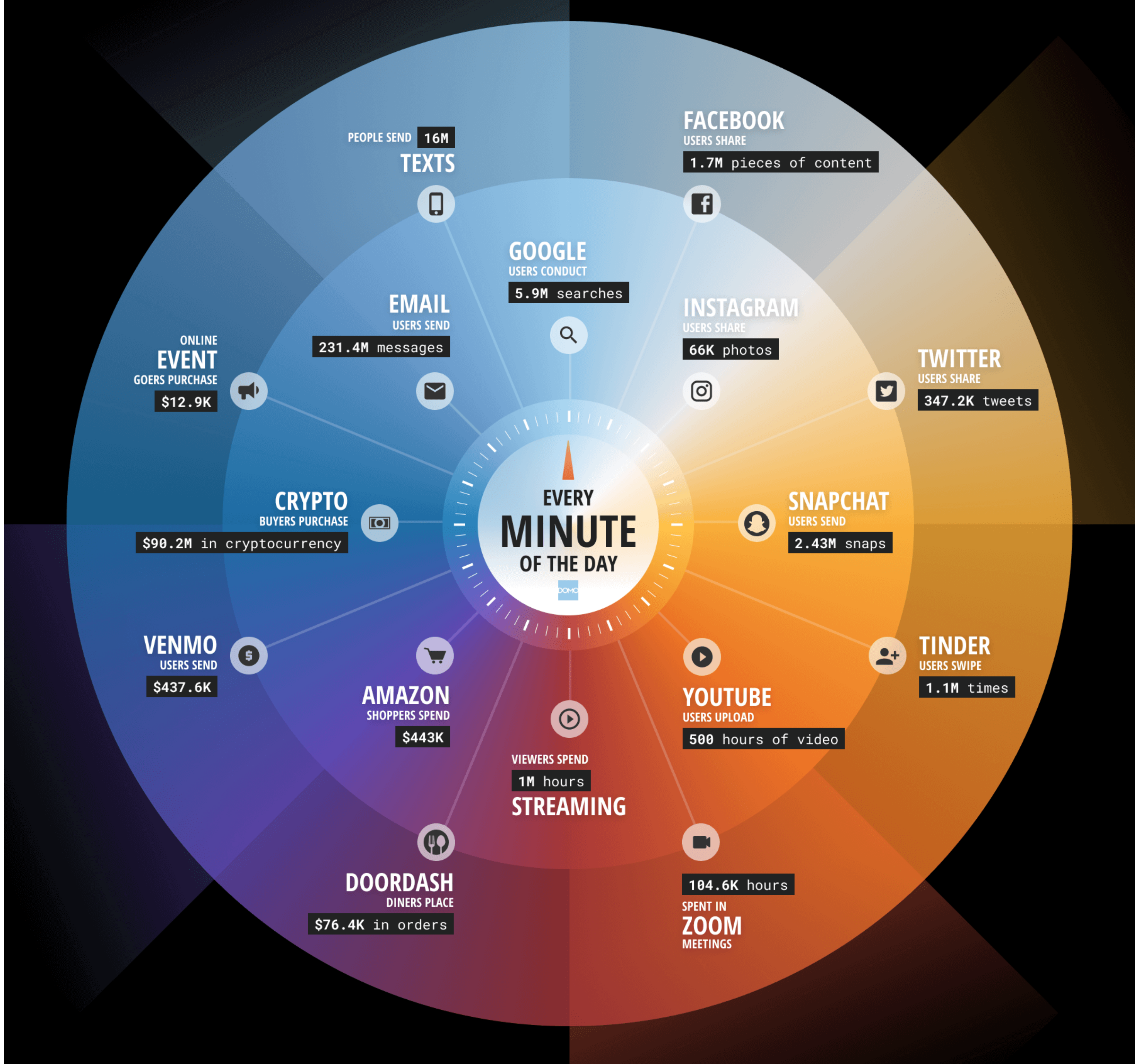
Biomedical Data Science:
Data Privacy



Eric Ni

CBB752b23

DATA NEVER SLEEPS 10.0



EVERY MINUTE OF THE DAY

GOOGLE USERS CONDUCT 5.9M searches

FACEBOOK USERS SHARE 1.7M pieces of content

INSTAGRAM USERS SHARE 66K photos

TWITTER USERS SHARE 347.2K tweets

SNAPCHAT USERS SEND 2.43M snaps

TINDER USERS SWIPE 1.1M times

YOUTUBE USERS UPLOAD 500 hours of video

104.6K hours SPENT IN **ZOOM** MEETINGS

VIEWERS SPEND 1M hours **STREAMING**

DOORDASH DINERS PLACE $76.4K in orders

AMAZON SHOPPERS SPEND $443K

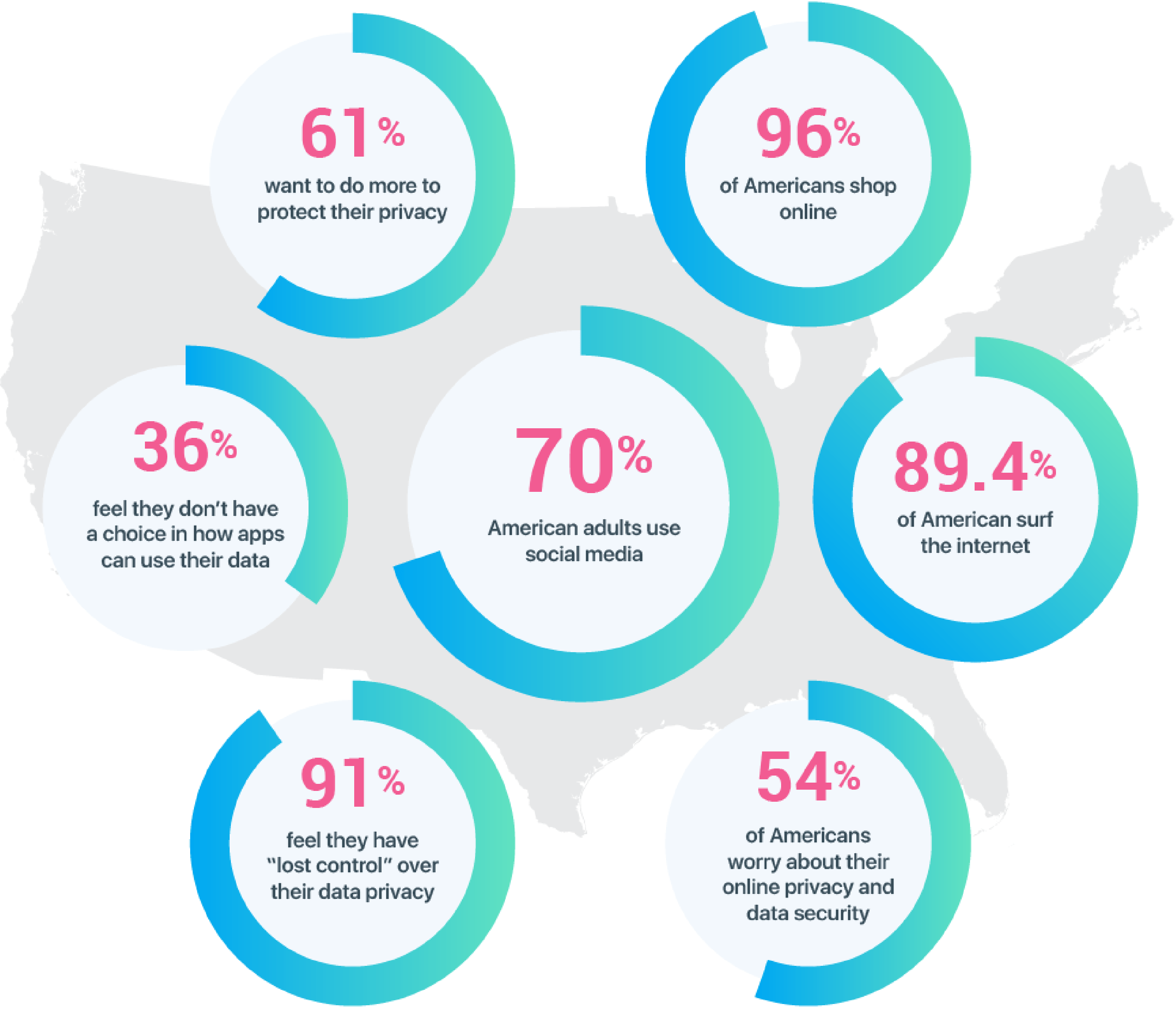
VENMO USERS SEND $437.6K

CRYPTO BUYERS PURCHASE $90.2M in cryptocurrency

ONLINE **EVENT** GOERS PURCHASE $12.9K

EMAIL USERS SEND 231.4M messages

PEOPLE SEND 16M **TEXTS**



61% want to do more to protect their privacy

96% of Americans shop online

36% feel they don't have a choice in how apps can use their data

70% American adults use social media

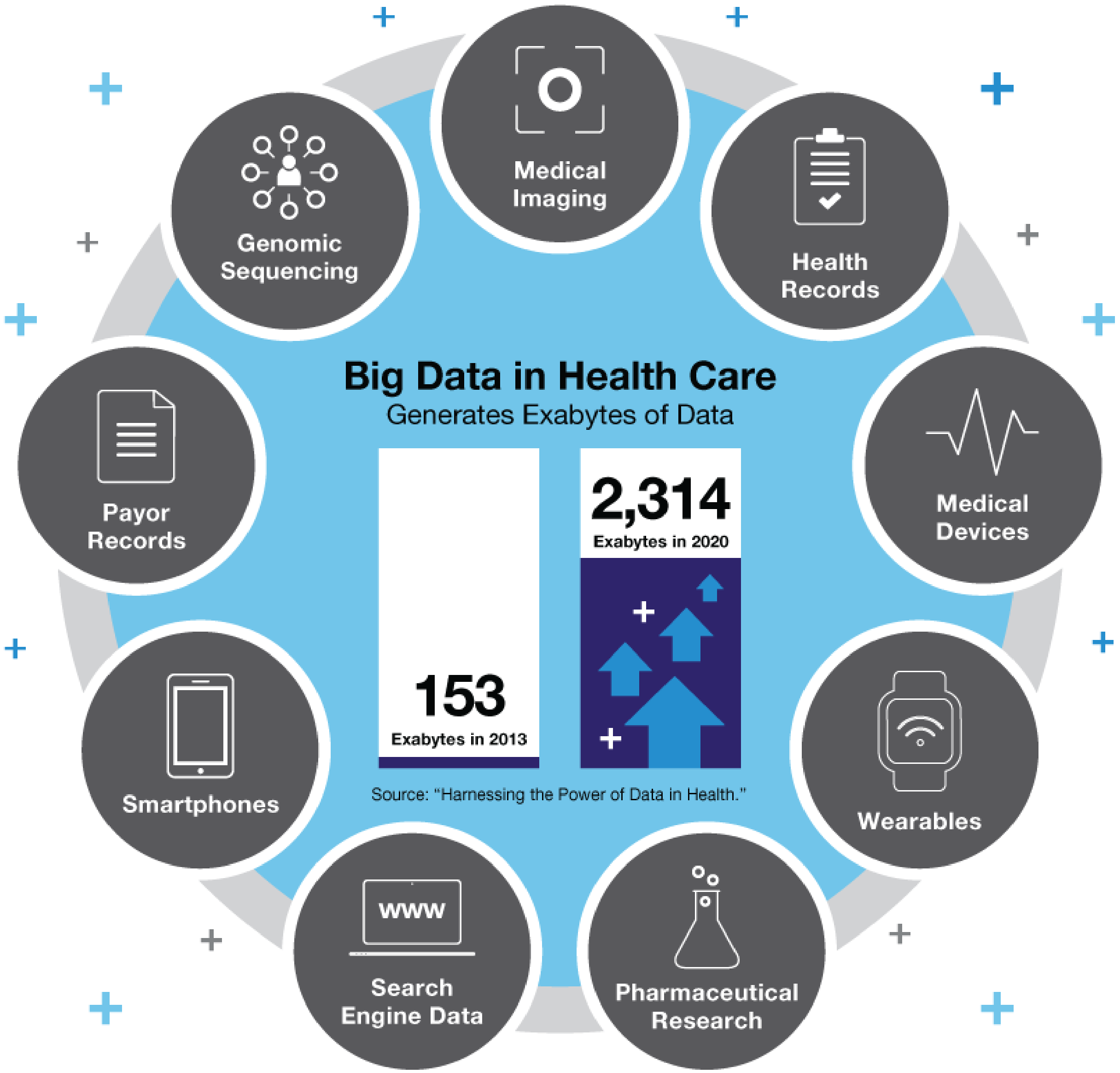
89.4% of American surf the internet

91% feel they have "lost control" over their data privacy

54% of Americans worry about their online privacy and data security

Sources: Pew Research Center, BigCommerce, Internet World Stats, Mobile Ecosystem Forum 2018, EpressVPN]

Where is all the health care data coming from?



Source: "Harnessing the Power of Data in Health."

What is privacy anyway?

- “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)
    - consent, the rights to be informed, of controlling/restricting access, of rectification, and erasure

Who owns your health data?

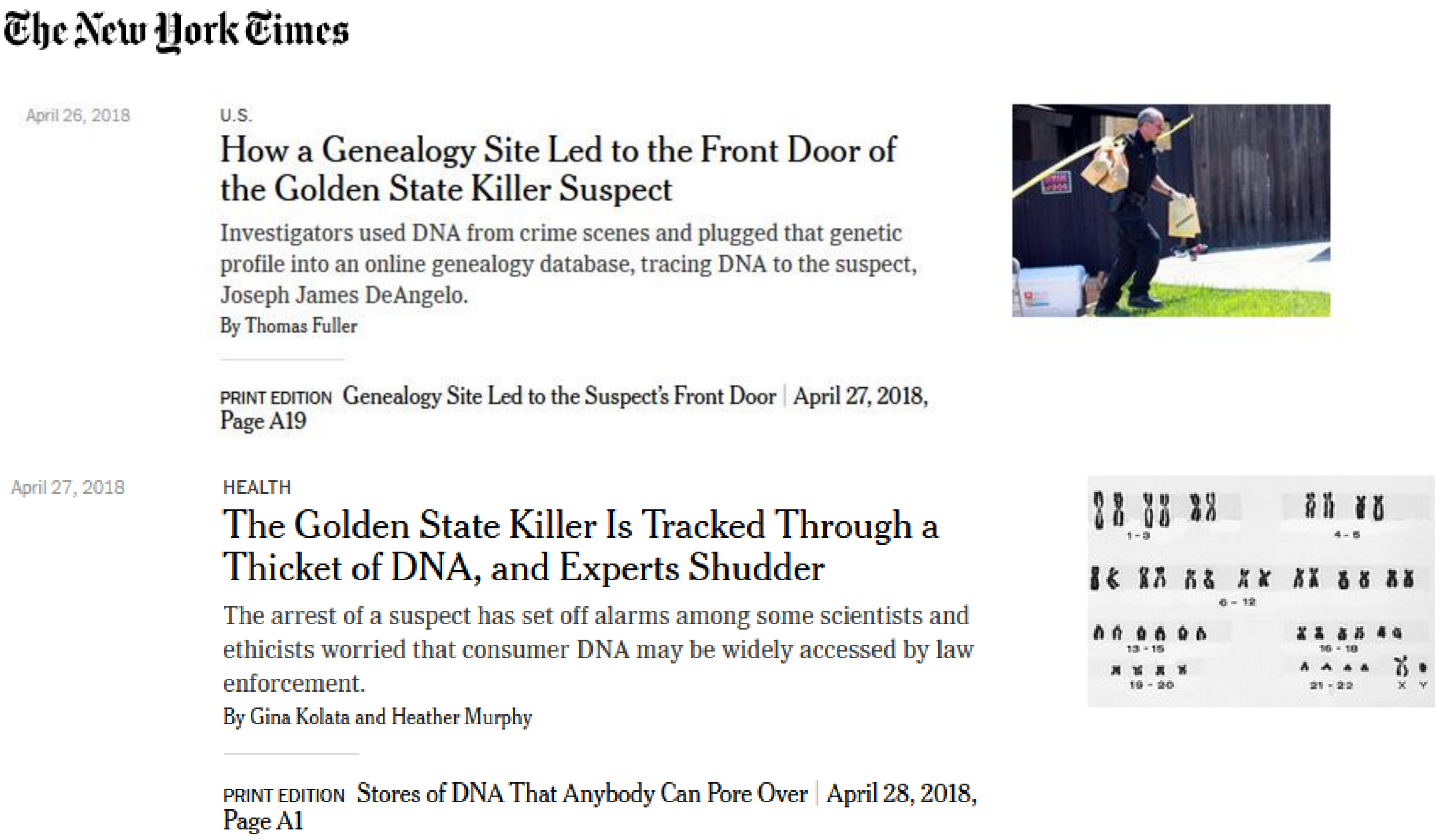
- Legally, varies by state, but usually, not the patients
  - In most states, legal ownership still resides in your healthcare provider
- HIPAA establishes standards for protecting “individually identifiable health information”, and patients can “inspect, review and receive a copy of his or her own medical records and billing records”

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

https://cphs.berkeley.edu/hipaa/hipaa18.html

Re-identification using genetic data



The New York Times

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

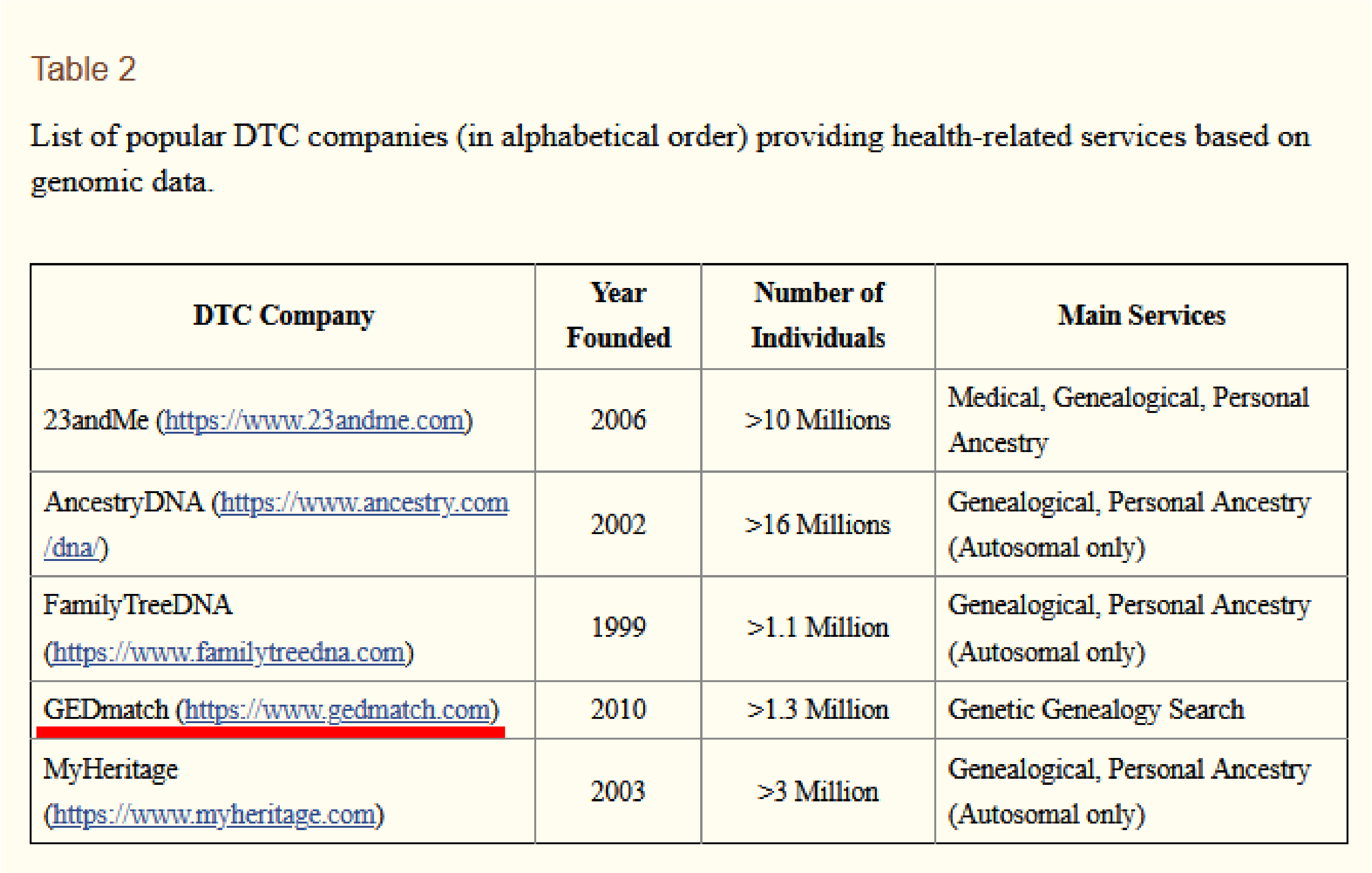
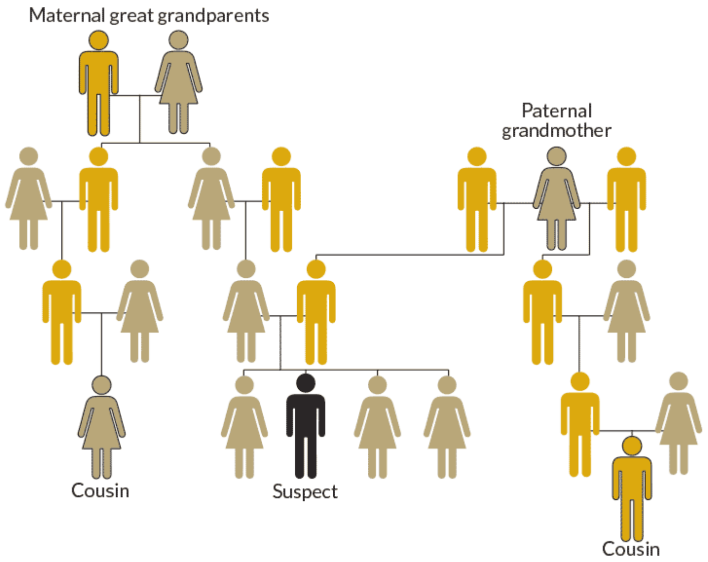


Table 2

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)
GEDmatch (https://www.gedmatch.com)	2010	>1.3 Million	Genetic Genealogy Search
MyHeritage (https://www.myheritage.com)	2003	>3 Million	Genealogical, Personal Ancestry (Autosomal only)

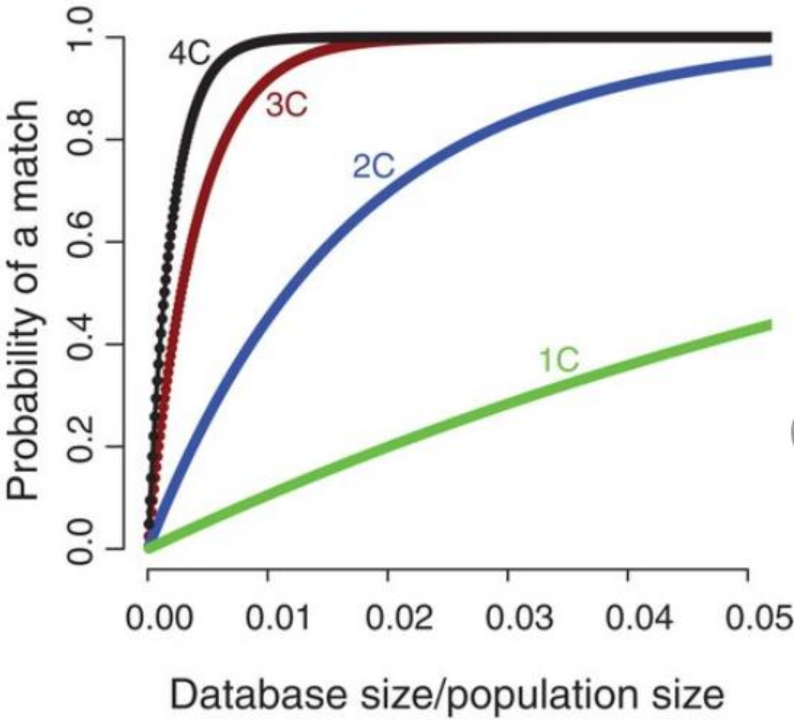


Bonomi et al. Nature Genetics (2020)
https://www.sciencenews.org/article/genetic-genealogy-forensics-top-science-stories-2018-yir

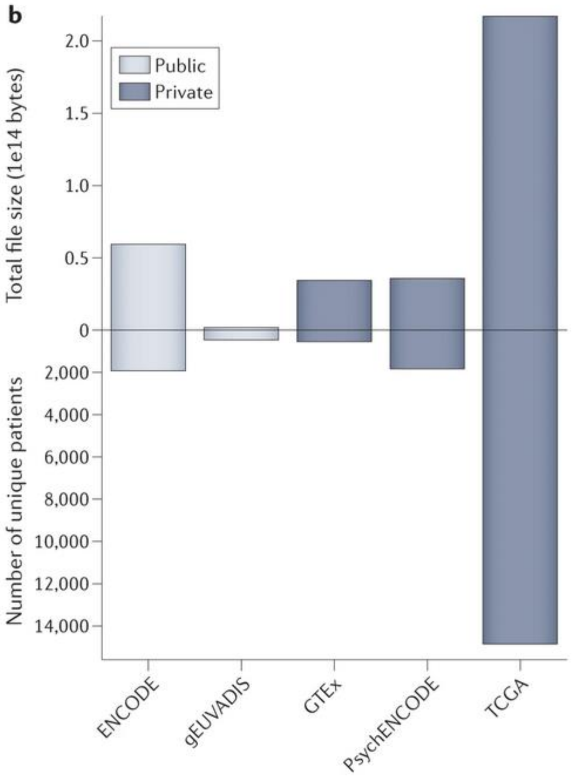
Identity inference of genomic data using long-range familial searches

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

SCIENCE • 11 Oct 2018 • Vol 362, Issue 6415 • pp. 690-694 • DOI: 10.1126/science.aau4832

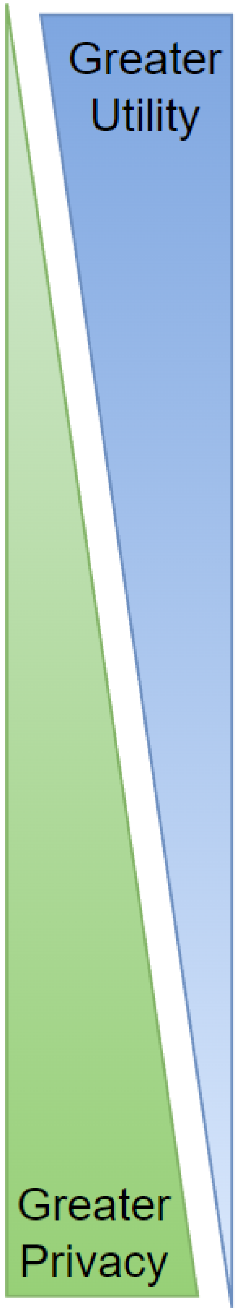


Biomedical data: To share … or not?

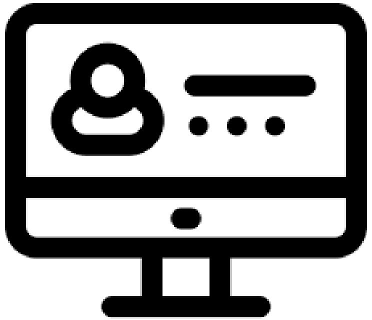


db GaP
GENOTYPES and PHENOTYPES

Privacy vs Utility



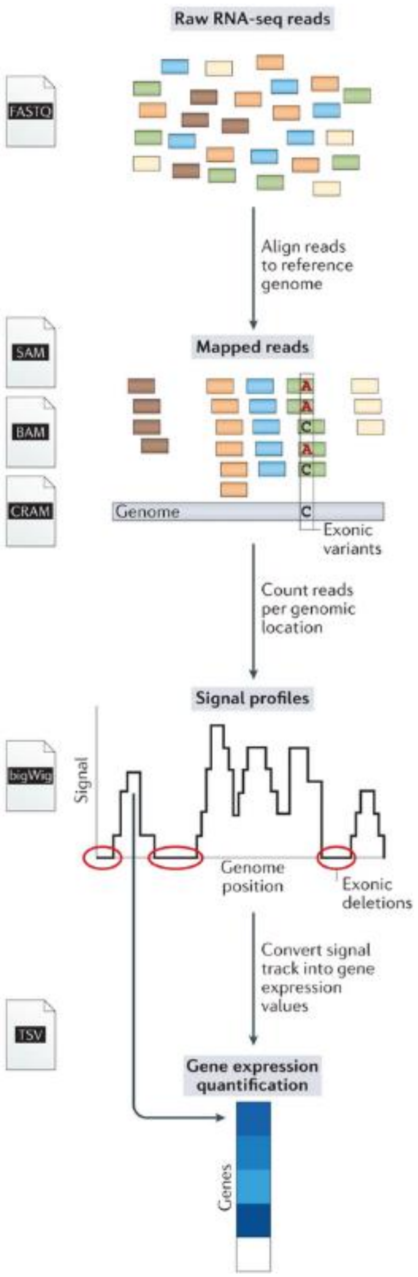
Open Access



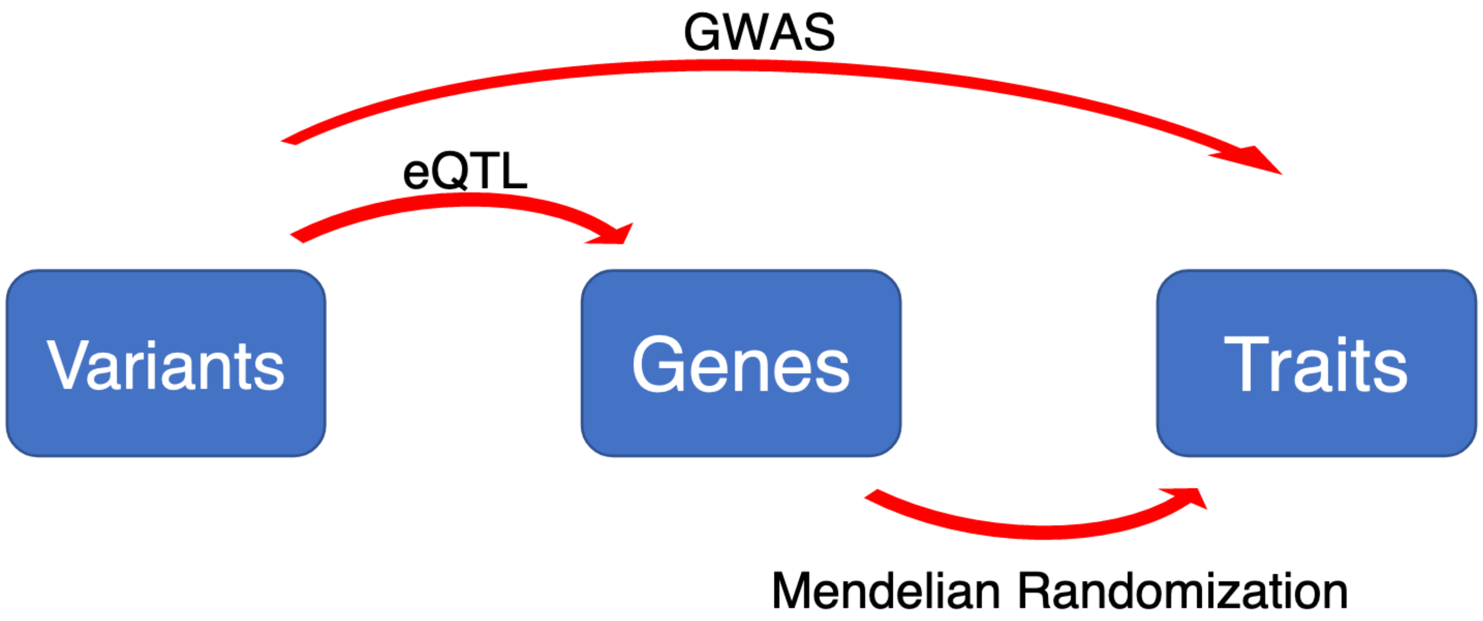
Registered Access

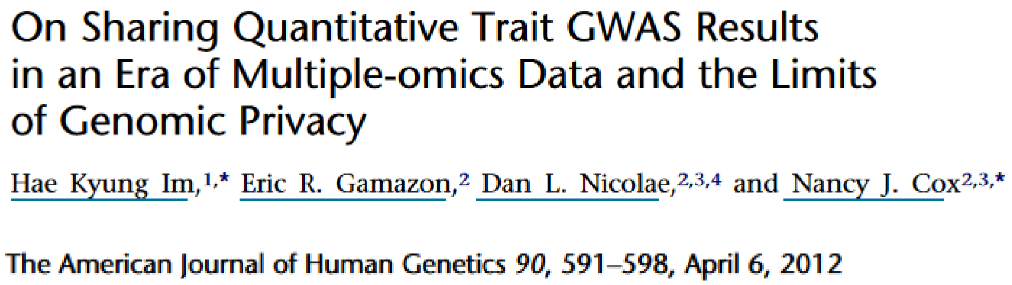
dbGaP GENOTYPES and PHENOTYPES

Controlled Access


Raw RNA-seq reads
FASTQ
Align reads to reference genome
Mapped reads
SAM
BAM
CRAM
A
A
C
A
C
Genome
C
Exonic variants
Count reads per genomic location
Signal profiles
bigWig
Signal
Genome position
Exonic deletions
Convert signal track into gene expression values
TSV
Gene expression quantification
Genes

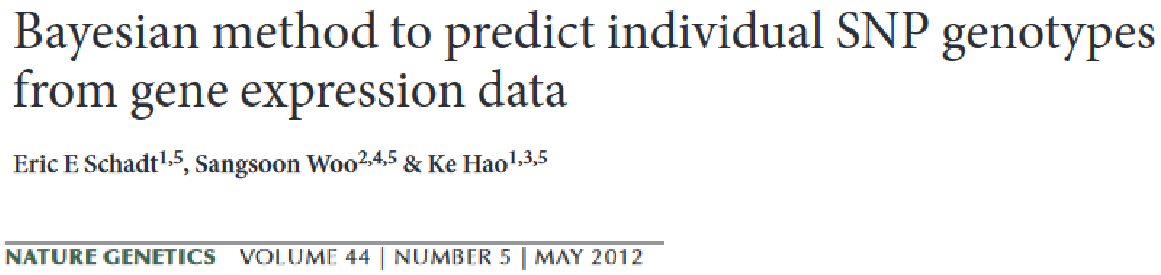
Privacy leakage in functional genomics




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

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
…	…	…	…
…	…	…	…

Names available for many users!

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

NETFLIX

IMDb

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
…	…	…	…
…	…	…	…

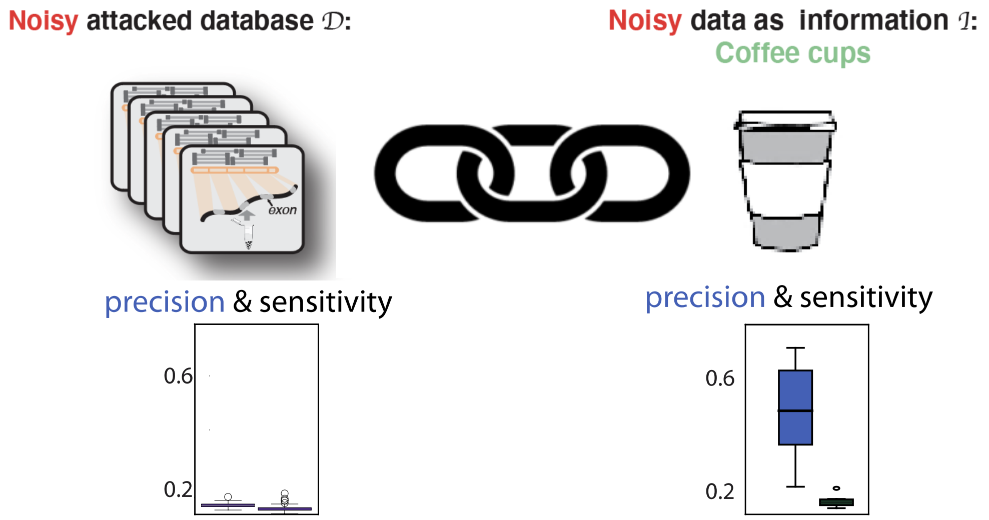
User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

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

- **IMDB users are public**

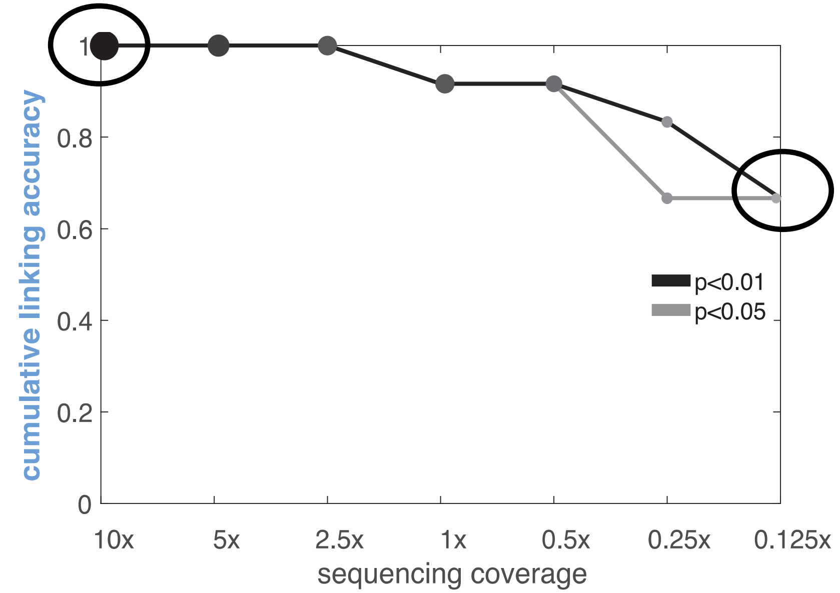
- **NetFLIX and IMdB moves are public**

Linking attack: genotype can be linked to reveal phenotypes



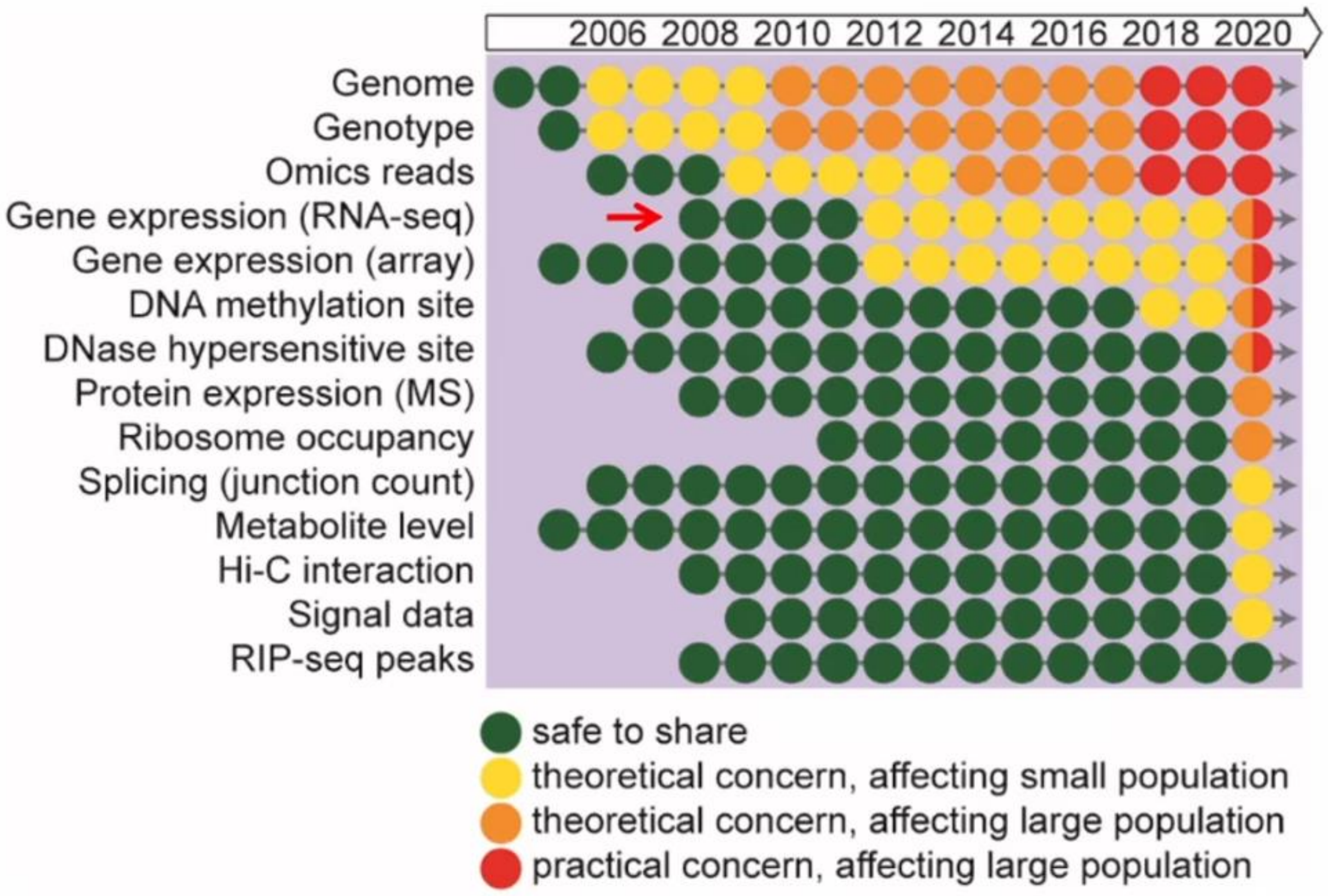
Noisy attacked database $\mathcal{D}$:

Noisy data as information $\mathcal{I}$: Coffee cups



Gursoy et al., **Cell**, 2020

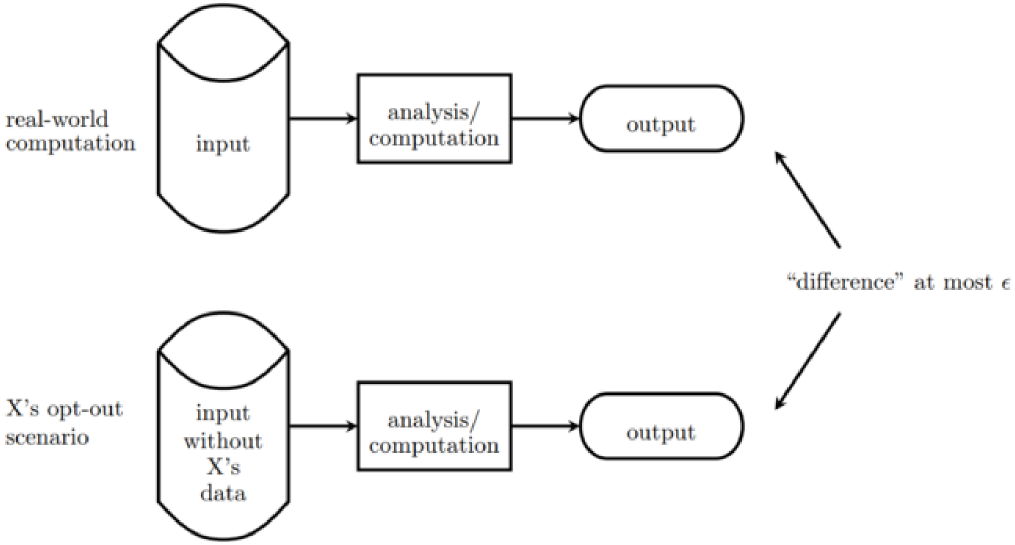
Latent functional risk in genomics data manifests over time

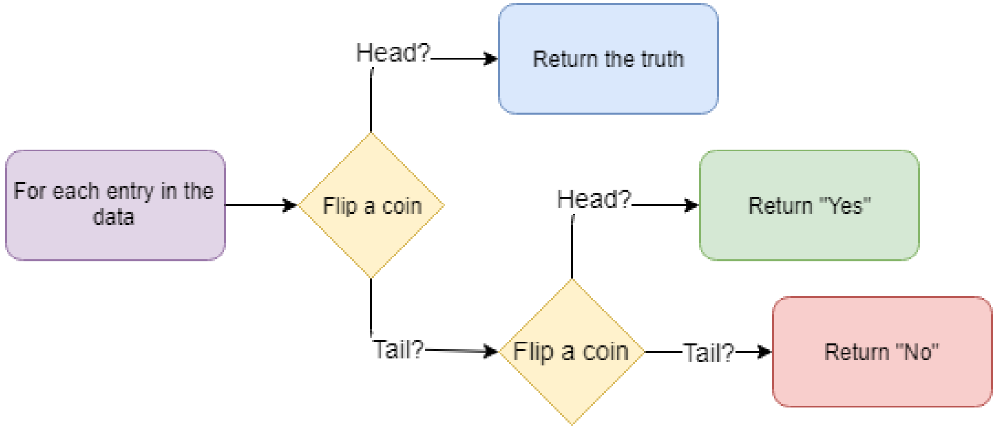


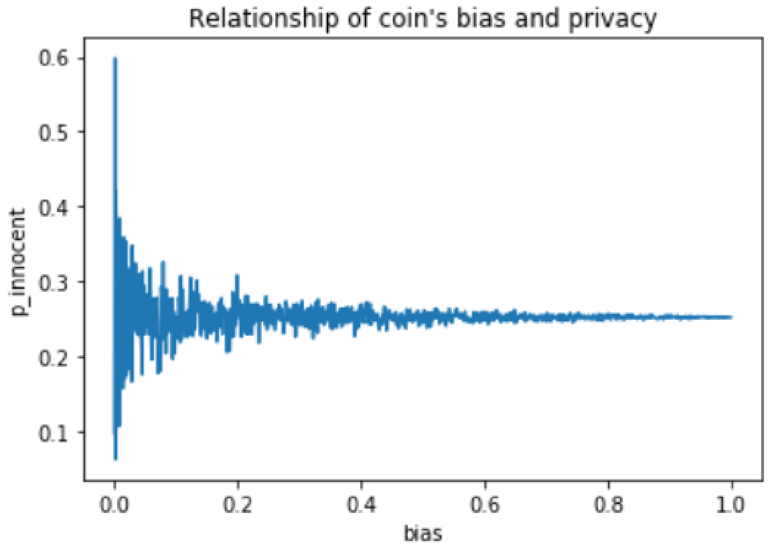
Slide from Zhiqiang Hu, GenoPri’21

Differential privacy

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

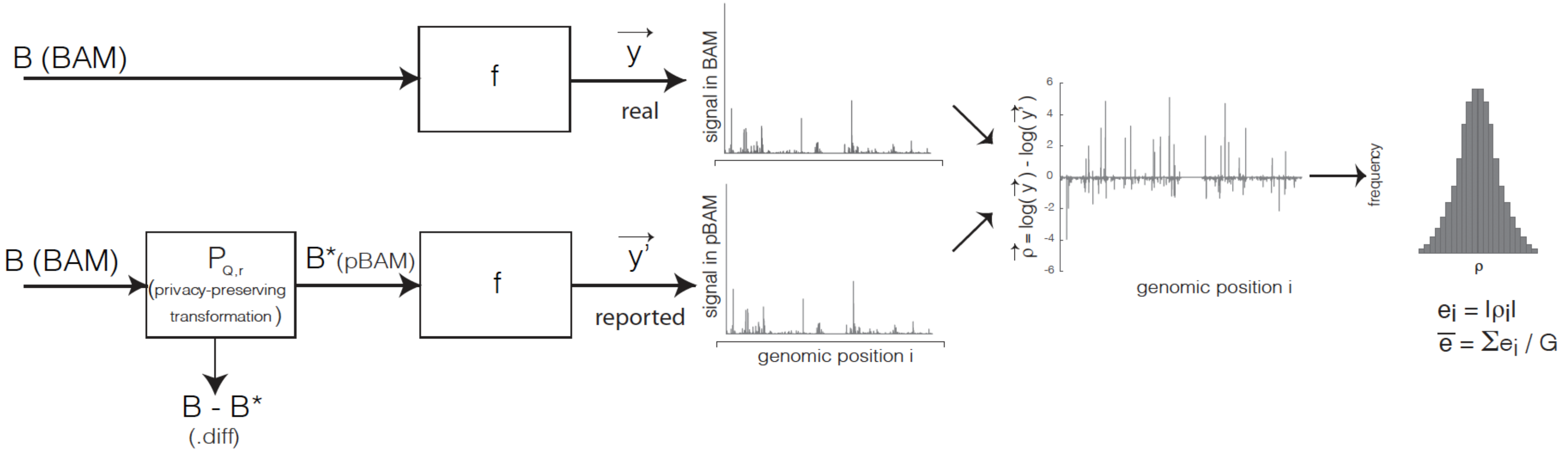




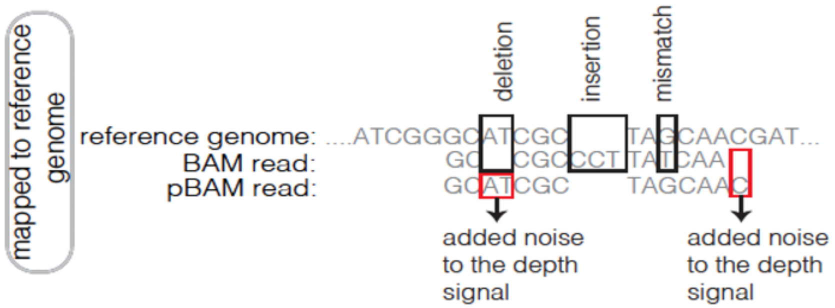


https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a

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



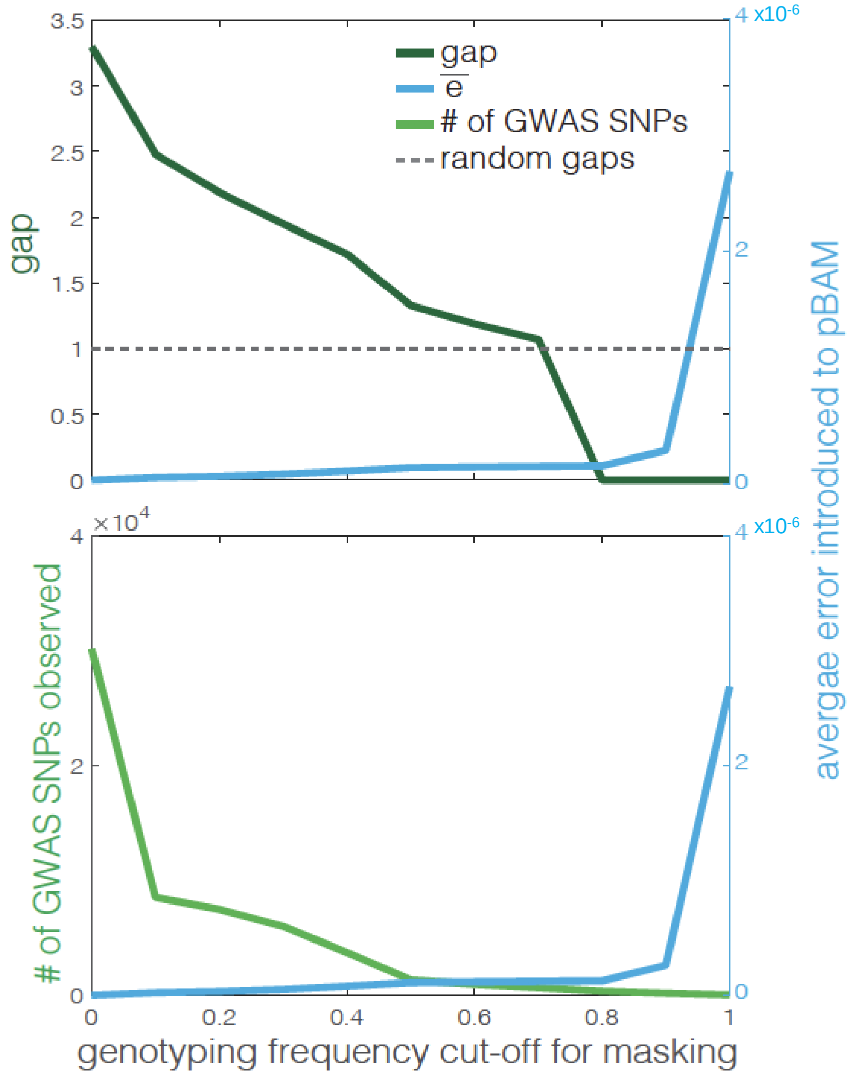
mapped to reference genome
deletion
insertion
mismatch
reference genome: ....ATCGGGCATCGC TAGCAACGAT...
BAM read: GC CGCCCT TATCAA
pBAM read: GCATCGC TAGCAAC
added noise to the depth signal
added noise to the depth signal

[Gursoy et al., Cell, 2020]

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



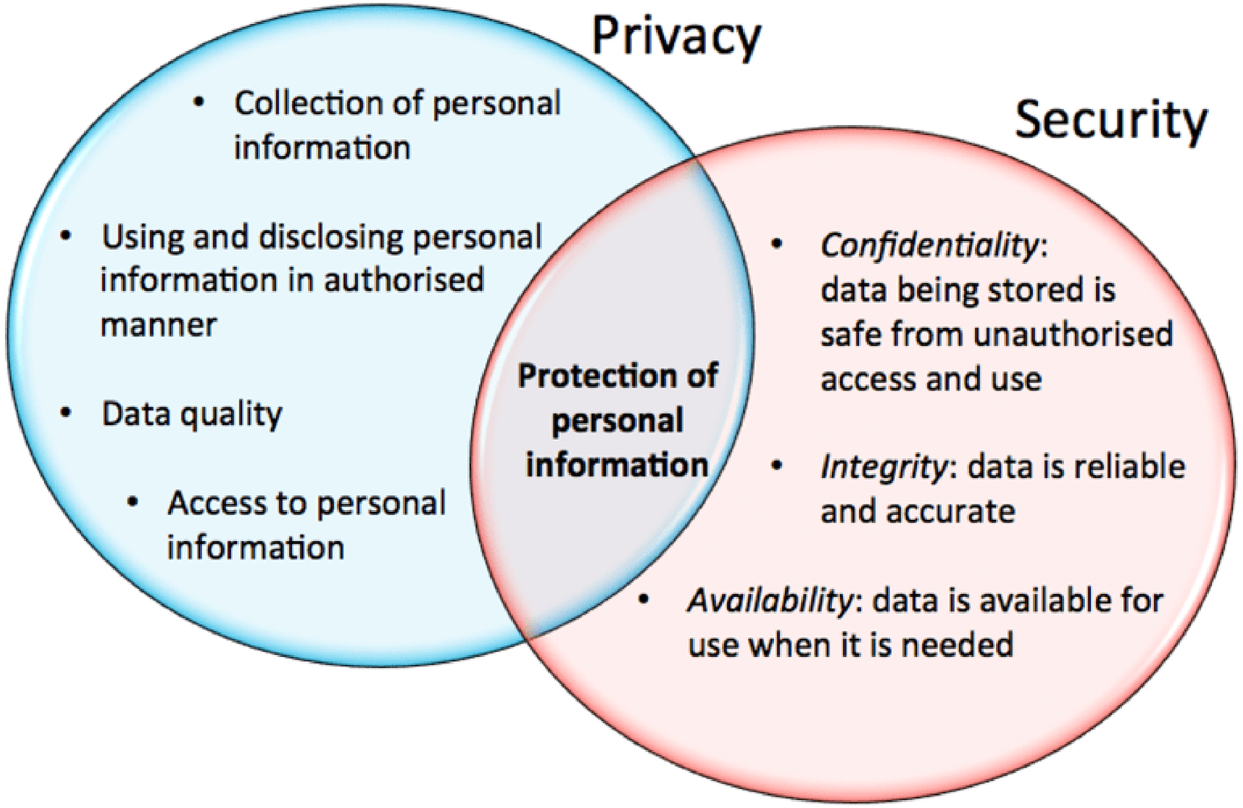
ε decreases as we introduce more error

δ increases as we mask more and more common variants

[Gursoy et al., Cell, 2020]

Privacy & Security

Privacy is different than Security



Privacy

- Collection of personal information
- Using and disclosing personal information in authorised manner
- Data quality
- Access to personal information

Protection of personal information

Security

- *Confidentiality*: data being stored is safe from unauthorised access and use
- *Integrity*: data is reliable and accurate
- *Availability*: data is available for use when it is needed

Figure: https://www.davidfroud.com/information-security-vs-privacy-are-the-lines-blurring/

Biomedical data storage needs

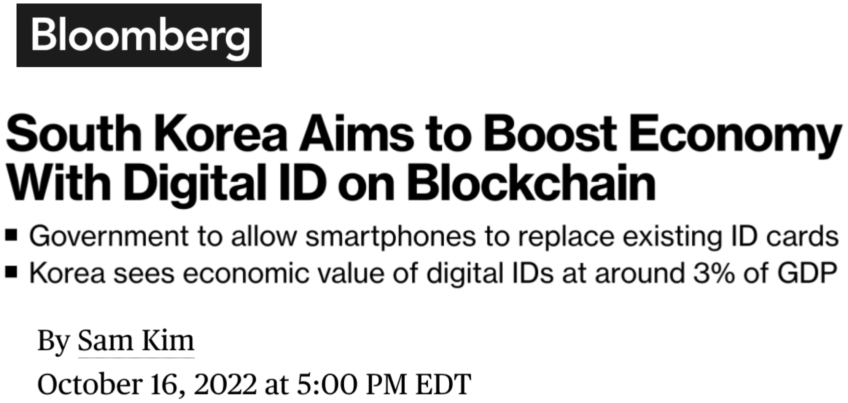
- **Data integrity**: ensuring accuracy and reliability for data during its entire life cycle
- **Access control**: appropriate access to those who need it, and not to those who don’t
- **Ownership rights**: ability to access, create, modify, package, derive benefit from, sell, or remove the data, and also the right to assign these access privileges to others

Blockchain can be useful for data storage/sharing

Why?

- **Decentralization** - information on a blockchain is distributed across a network of computers, prevents a single point of failure
- **Immutability** - once data is added to the blockchain, it cannot be altered or removed.
- **Auditability** - the ability to easily track and verify the history of the blockchain

Blockchain has many potential non-financial applications

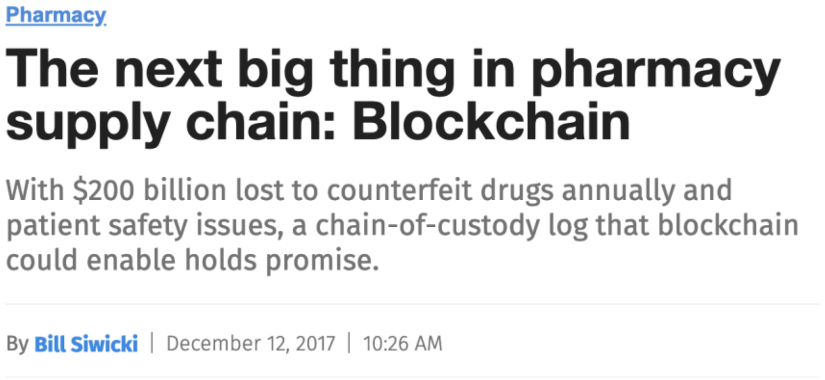


Bloomberg

South Korea Aims to Boost Economy With Digital ID on Blockchain

- Government to allow smartphones to replace existing ID cards
- Korea sees economic value of digital IDs at around 3% of GDP

By Sam Kim
October 16, 2022 at 5:00 PM EDT

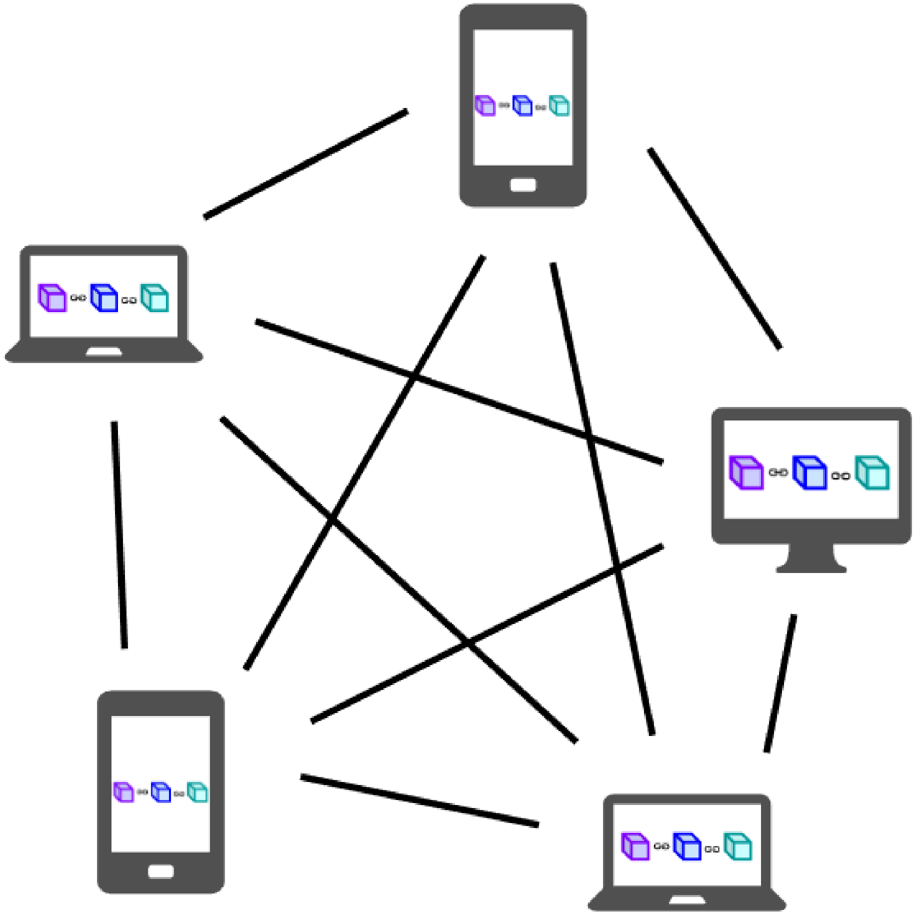


Pharmacy

The next big thing in pharmacy supply chain: Blockchain

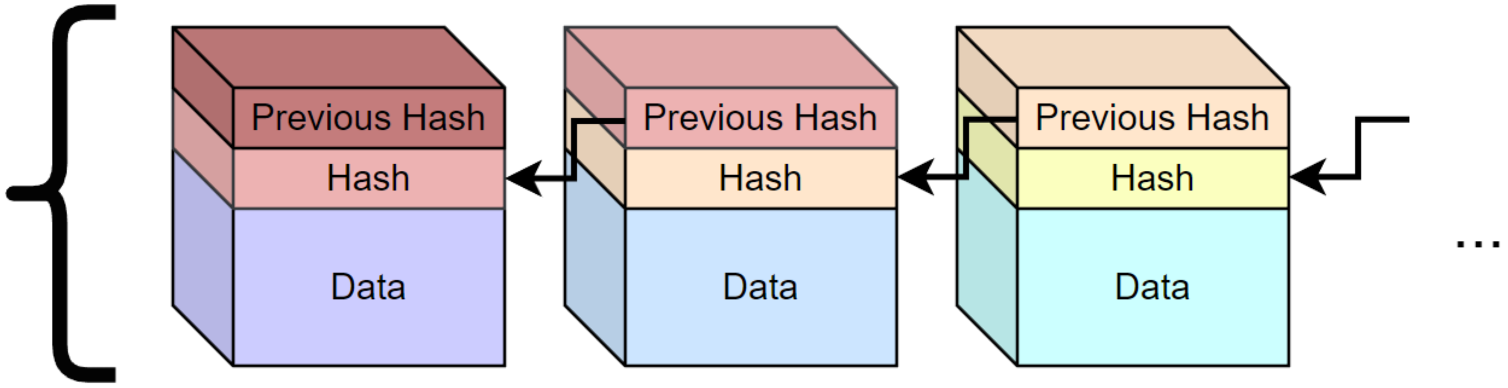
With $200 billion lost to counterfeit drugs annually and patient safety issues, a chain-of-custody log that blockchain could enable holds promise.

By Bill Siwicki | December 12, 2017 | 10:26 AM



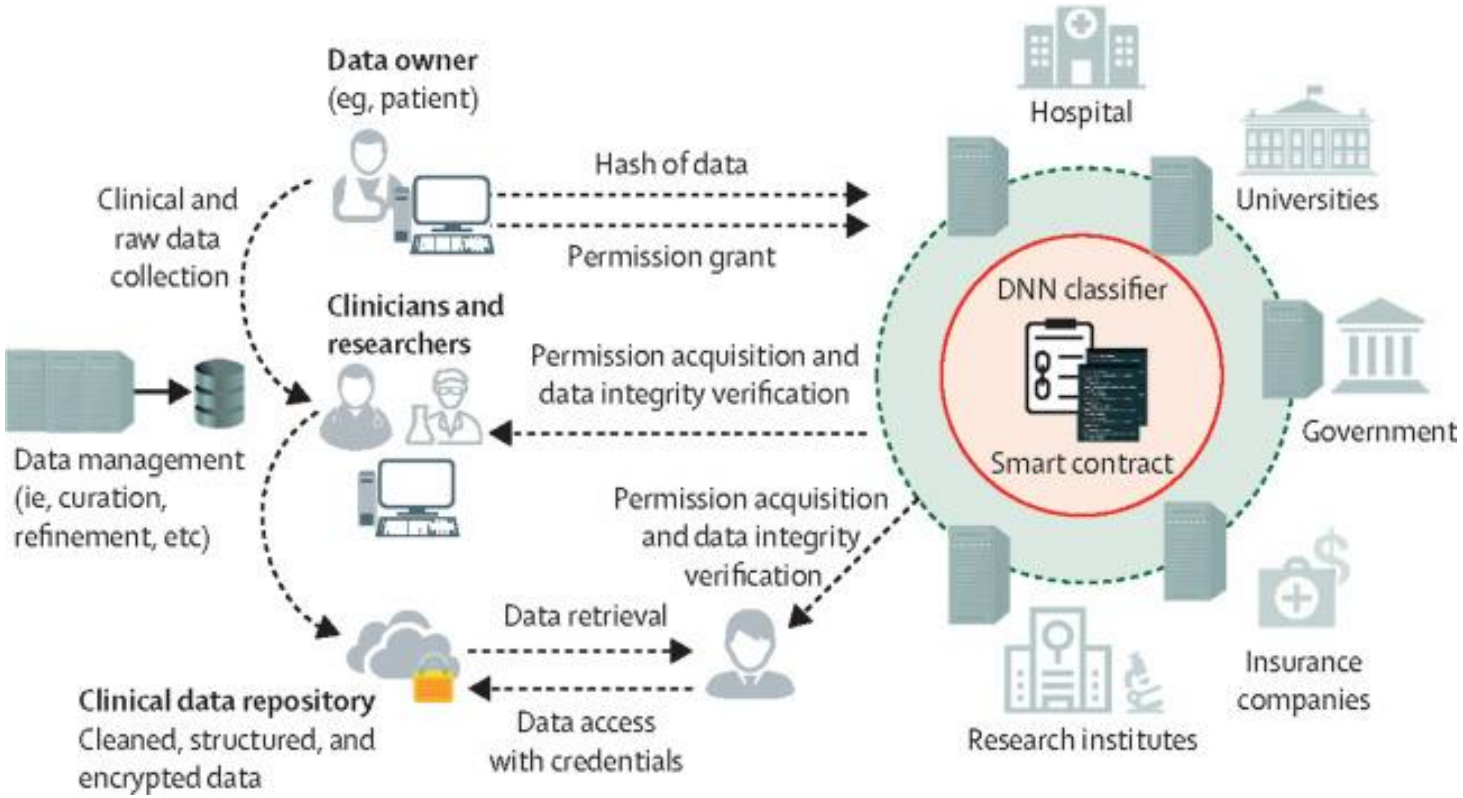
Blockchain:

- **Distributed** ledgers of information
- **Synchronized** across all participants
- Cryptographic links for data **immutability**



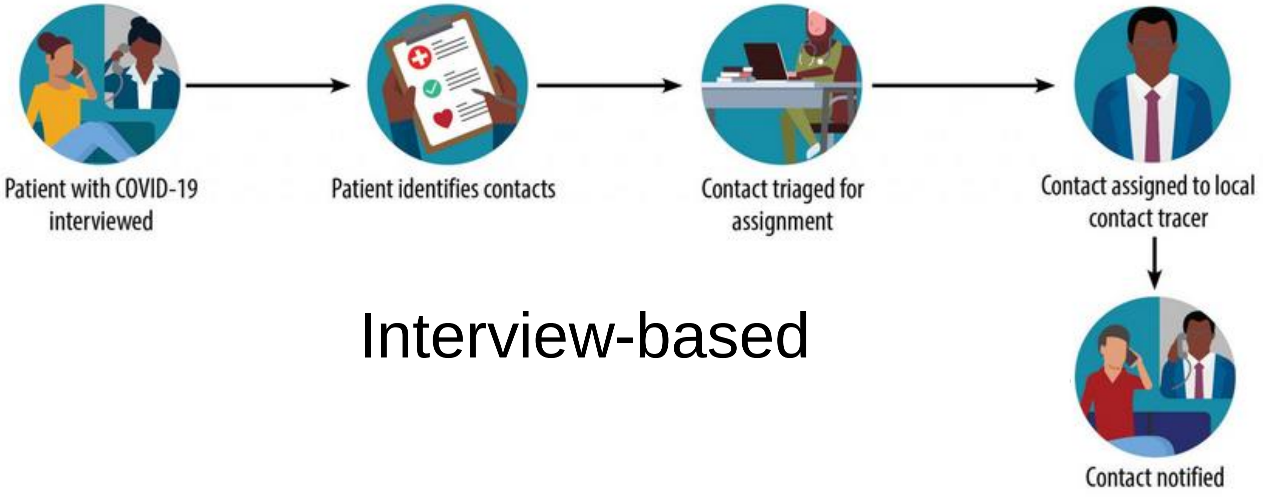
What is blockchain?

Blockchain: a solution for EHR sharing



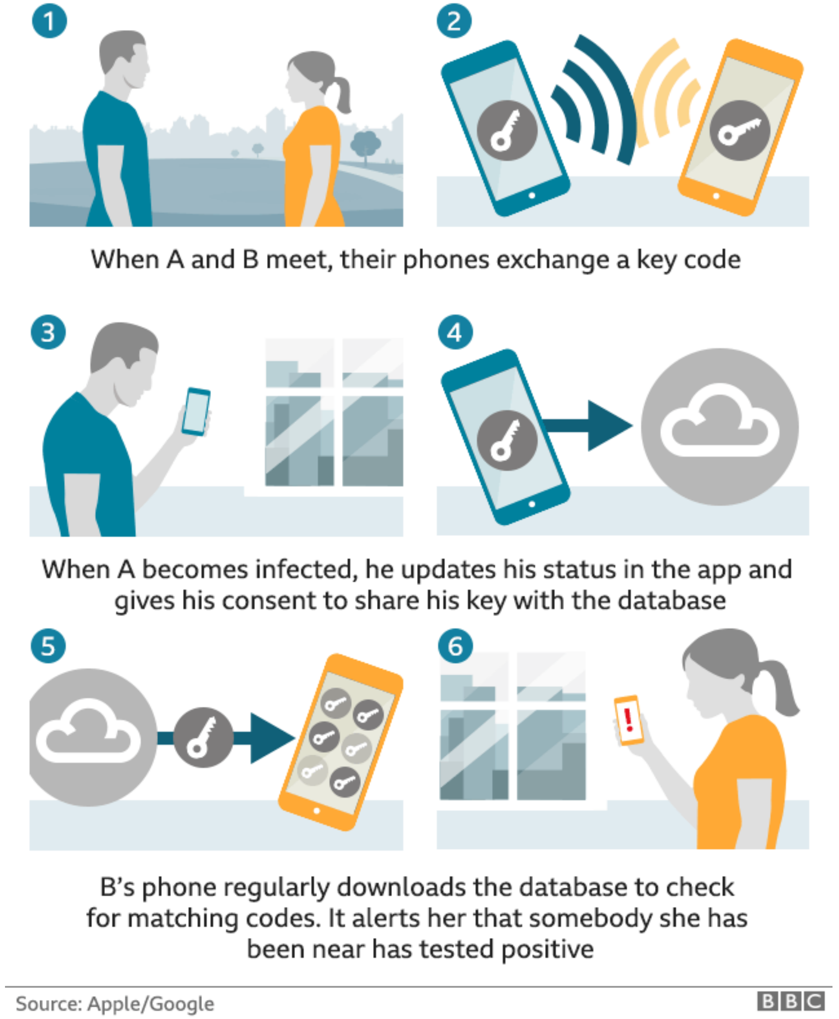
Ng et al. Lancet Digital Health (2021)

Contact tracing


Patient with COVID-19 interviewed
Patient identifies contacts
Contact triaged for assignment
Contact assigned to local contact tracer
Contact notified

Interview-based

App-based



When A and B meet, their phones exchange a key code

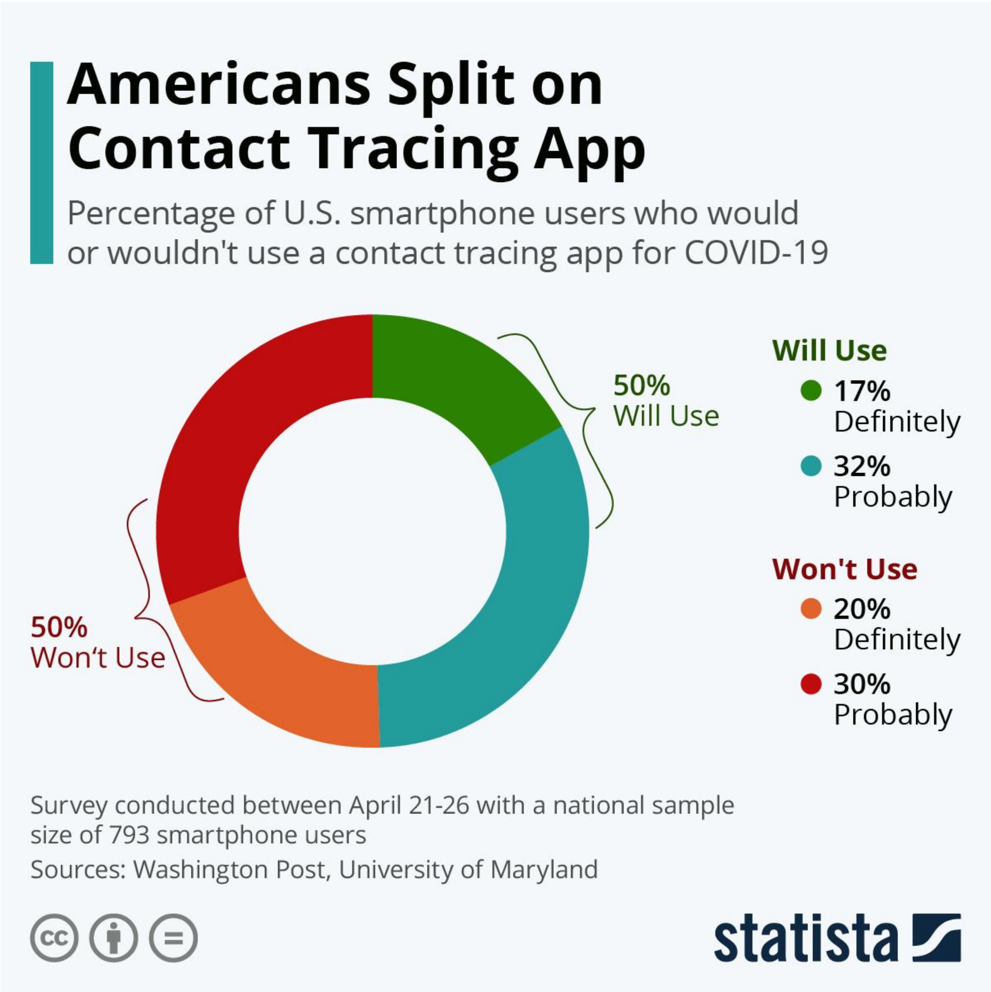
When A becomes infected, he updates his status in the app and gives his consent to share his key with the database

B's phone regularly downloads the database to check for matching codes. It alerts her that somebody she has been near has tested positive

Source: Apple/Google

https://www.cdc.gov/coronavirus/2019-ncov/global-covid-19/contact-tracing-workflow.html

Utility practically

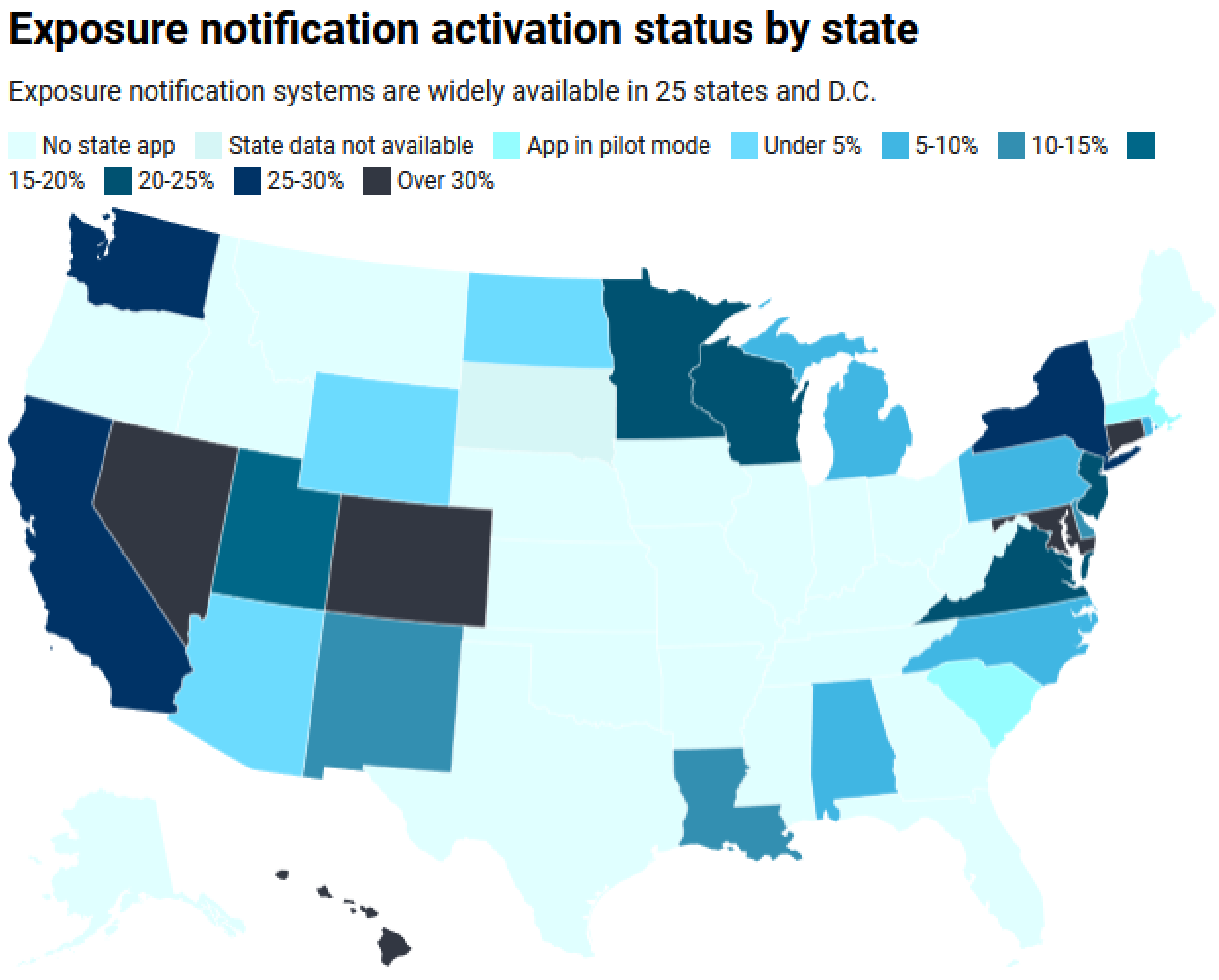


Americans Split on Contact Tracing App

Percentage of U.S. smartphone users who would or wouldn't use a contact tracing app for COVID-19

Survey conducted between April 21-26 with a national sample size of 793 smartphone users

Sources: Washington Post, University of Maryland



Exposure notification activation status by state

Exposure notification systems are widely available in 25 states and D.C.

North Dakota's figure represents active users, not total downloads. For D.C., the broader metro area population was used (rather than District residents only) because anyone living/working in D.C. may use this EN system.

Map: Betsy Ladyzhets / MIT Technology Review • Source: State public health departments, US Census • Get the data • Created