November 8, 2023

De-identification Workshop

Ann Waldo, JD
Principal
Waldo Law Offices

Daniel Barth-Jones, PhD
Privacy Expert in Residence
Privacy Hub by Datavant

Claire Manneh, MPH
Head of Provider Research
Datavant

Andrew Kopelman, JD
SVP, DGC, & Chief Privacy Counsel
Medidata

Privacy+
Security
Forum

Speaker

Ann Waldo, JD
Principal,
Waldo Law Offices, PLLC

Ann Waldo is the Principal in the boutique law firm of Waldo Law Offices in Washington, DC. She provides legal counsel regarding health data privacy, data strategy, and data transactions, as well as public policy and advocacy regarding data privacy. She has worked as Chief Privacy Officer for Lenovo, Chief Privacy Officer at Hoffmann-La Roche, in Public Policy at GlaxoSmithKline, in-house counsel at IBM, and commercial litigation. Ann has a JD from UNC Law School with high honors. She is licensed to practice law in DC and North Carolina and is a member of the Bar of the U.S. Supreme Court. She is passionate about health data and innovation.

Speaker

Privacy+ Security Forum

Daniel Barth-Jones, MPH, PhD
Privacy Expert in Residence
Privacy Hub by Datavant

Dr. Barth-Jones has conducted and managed statistical disclosure limitation operations and research involving activities in the healthcare information industry and in academia for more than two decades. His focus has been how to best balance protections for the privacy of individuals within health information databases while simultaneously preserving the analytic accuracy of statistical analyses. He has provided educational training and made numerous scientific presentations on statistical disclosure limitation to federal agencies, national and state healthcare organizations, commercial healthcare/healthcare information companies, and in academia. He joined *Privacy Hub by Datavant* in June of 2022 as a Principal Privacy Expert. Prior to joining Privacy Hub, Dr. Barth-Jones was an Assistant Professor of Clinical Epidemiology on the faculty of the Department of Epidemiology at Columbia University from 2007 to 2022 and was a faculty member in the Center for Healthcare Effectiveness Research at the Wayne State University Medical School from 2000 to 2006. Daniel was also the Founder and President of dEpid/dt Consulting for more than twenty years. He received his Master of Public Health degree in General Epidemiology and Ph.D. in Epidemiologic Science from the University of Michigan.

Speaker

Privacy+ Security Forum

Claire Manneh, MPH
Head of Provider Partnerships
Datavant

Claire is the Head of Provider Partnerships for Datavant and works closely with academic medical centers, health systems, and research collaboratives. Previously, Claire led provider partnerships at Included Health and spent several years at the California Hospital Association as the Director of the California Hospital Patient Safety Organization. Claire was a board member on the California Maternal Quality Care Collaborative and liaised with the state's 240 birthing hospitals to support efforts in reducing unnecessary c-sections. As a Fulbright Scholar in the Sultanate of Oman, she studied the use of disparate electronic medical records across the country's three major hospitals and consulted on the need to move toward an NHS-like system with the Ministry of Health. Claire has a double B.A. in Political Science and Public Health from the University of California at Berkeley and a Master of Public Health from Dartmouth.

Speaker

Andrew Kopelman, JD
Senior Vice President, Deputy General Counsel & Chief Privacy Counsel
Medidata Solutions

Andrew's focus is on the intersection of SaaS technology, health data innovation, and data protection. He heads Medidata's transactional, compliance and corporate functions related to data protection and IP, and is the lead attorney for Medidata AI, a division of hundreds of multi-disciplinary experts who deliver innovative solutions to the Life Sciences industry. Prior to joining Medidata in 2012, Andrew was an associate in the patent group at Jenner & Block (practicing patent litigation and counseling on pharmaceutical and computer patents) and at Frommer, Lawrence & Haug (engaged in Hatch-Waxman pharmaceutical litigation). Before law school, he was a computer programmer for technology startups and incubators. He is a Certified Information Privacy Professional (CIPP/US, CIPP/E) and a registered patent attorney. Andrew earned his B.A. with a concentration in Asian Studies from Carleton College and his J.D. as an associate editor of the Cardozo Law Review with a concentration in IP from the Cardozo School of Law. Andrew currently chairs the Global Data Protection and Privacy (GDPP) committee of ACRO (the Association of Contract Research Organizations).

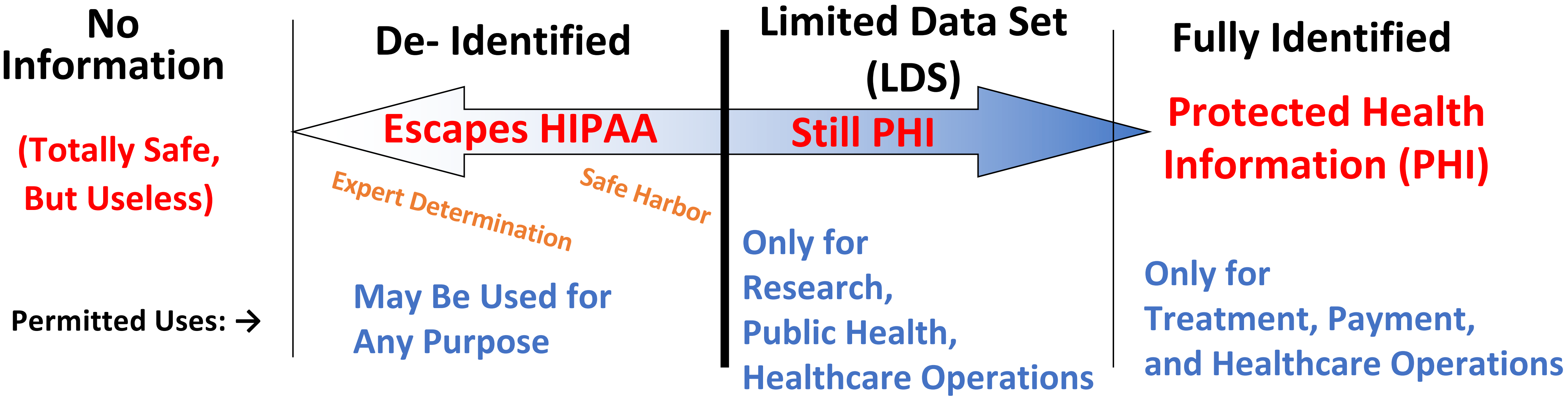
Overall Workshop Questions

- What is de-identification – under HIPAA, EU law, and evolving state laws?
- What are the statistical, technical, and privacy-preserving challenges?
- Why does de-identification matter in the real world? What can de-identified data accomplish?
- What's happening already with de-identified data that wasn't happening a few years ago?
- What new technologies can make it more viable to extract scientific insights from linked de-identified data ?
- How might AI affect de-identification?
- How have the new de-ID'n definitions in the new state laws changed things?
- What new state law obligations attach to de-ID'd data?
- How can the data ecosystem deal with the challenging and fast-changing de-ID'n environment?
- Are there reasons to hope for clarity around anonymisation under GDPR?

Framing De-Identification

Daniel Barth-Jones

HIPAA’s Identification Risk/Legal Spectrum



No Information

(Totally Safe, But Useless)

Permitted Uses: →

De- Identified

Escapes HIPAA

Expert Determination

Safe Harbor

May Be Used for Any Purpose

Limited Data Set (LDS)

Still PHI

Only for Research, Public Health, Healthcare Operations

Fully Identified

Protected Health Information (PHI)

Only for Treatment, Payment, and Healthcare Operations

Limited Data Set (LDS) §164.514(e)

Eliminate 16 Direct Identifiers (Name, Address, SSN, etc.)

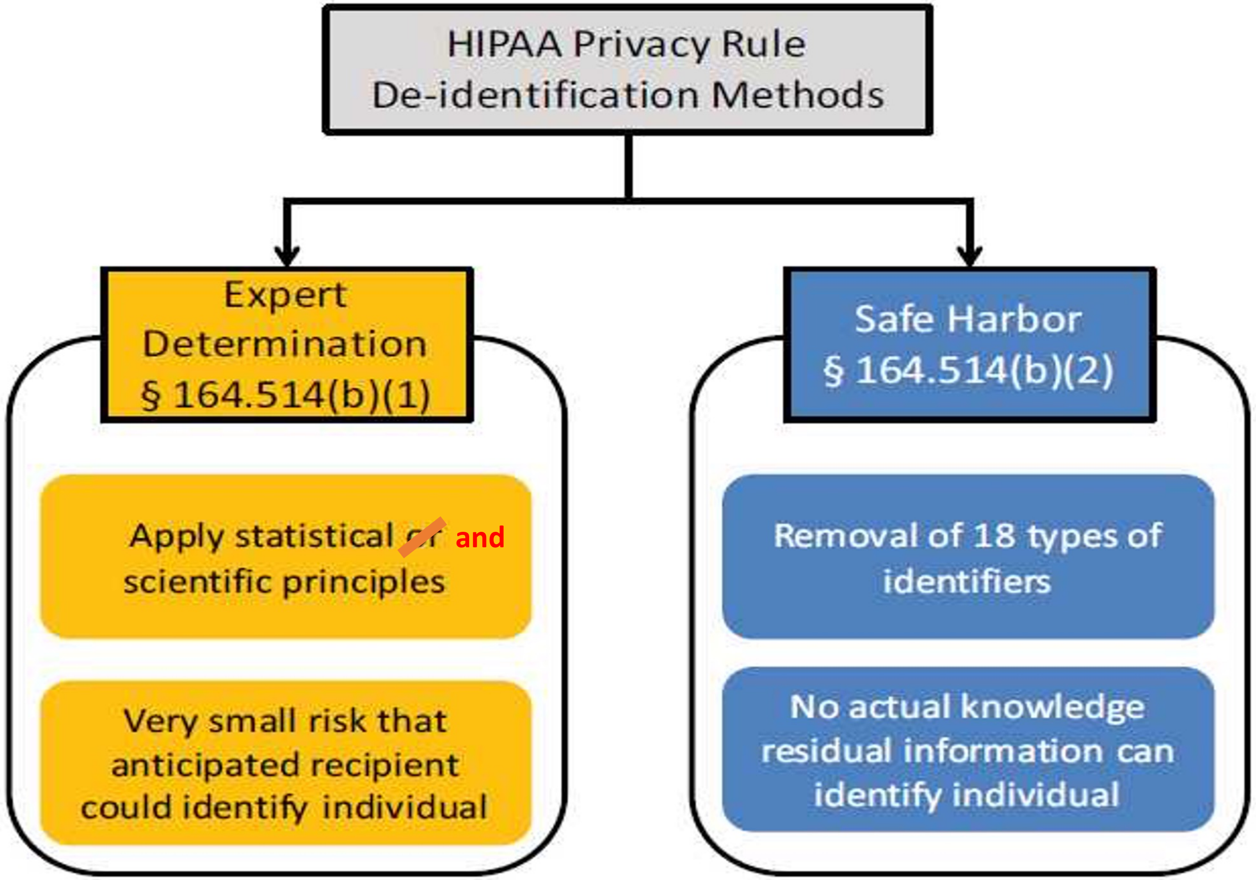
Safe Harbor De-identified §164.514(b)(2)

Eliminate 18 Identifiers (including Geography < 3-digit ZIP Code, and All Dates, except the Year)

Expert Determination Data Set (EDDS) §164.514(b)(1)

Expert’s Analysis Confirms a “Very Small” Risk of Re-identification

Two Methods of HIPAA De-identification



HIPAA Privacy Rule
De-identification Methods

Expert Determination
§ 164.514(b)(1)

Apply statistical ~~or~~ and scientific principles

Very small risk that anticipated recipient could identify individual

Safe Harbor
§ 164.514(b)(2)

Removal of 18 types of identifiers

No actual knowledge residual information can identify individual

Source: HHS Office for Civil Rights (OCR) De-Identification Guidance (November 2012)
[Corrected to match wording of §164.514(b)(1)]

HIPAA §164.514(b)(2)(i) -18 “Safe Harbor” Exclusions

All of the following must be **removed in order** for the information **to be** considered **de-identified**.

(2)(i) The **following identifiers of the individual or of relatives, employers, or household members** of the individual, are removed:

(A) Names;
(B) All **geographic subdivisions smaller than a State**, including street address, city, county, precinct, zip code, and their equivalent geocodes, **except for the initial three digits of a zip code** if, according to the current publicly available data from the Bureau of the Census: (*1*) The geographic unit formed by combining all zip codes with the same three initial digits contains **more than 20,000 people**; and (*2*) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
(C) **All elements of dates (except year)** for dates directly related to an individual, including **birth date**, **admission date**, **discharge date, date of death**; and **all ages over 89** and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) **Medical record numbers**;

(I) **Health plan beneficiary numbers**;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) **Device identifiers and serial numbers**;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and

(R) **Any other unique identifying number, characteristic, or code** except as permitted in §164.514(c)

Limits of Safe Harbor De-identification

- Full Dates and detailed Geography are often critical
- Challenging in complex data sets
  - Safe Harbor rules prohibiting Unique codes (§164.514(2)(i)(R)) unless they are not "derived from or related to information about the individual"(§164.514(c)(1)) can create significant complications for:
    - Preserving referential integrity in relational databases
    - Creating longitudinal de-identified data across parties
- Encryption does not equal de-identification
  - Encryption of PHI, rather than its removal - as required under safe harbor, will not necessarily result in de-identification
- Not convenient for "Data Masking"
  - Removal requirement in 164.514(b)(2)(i)
  - Software development requires realistic "fake" data which can pose re-identification risks if not properly managed

HIPAA §164.514(b)(1) “Expert Determination”

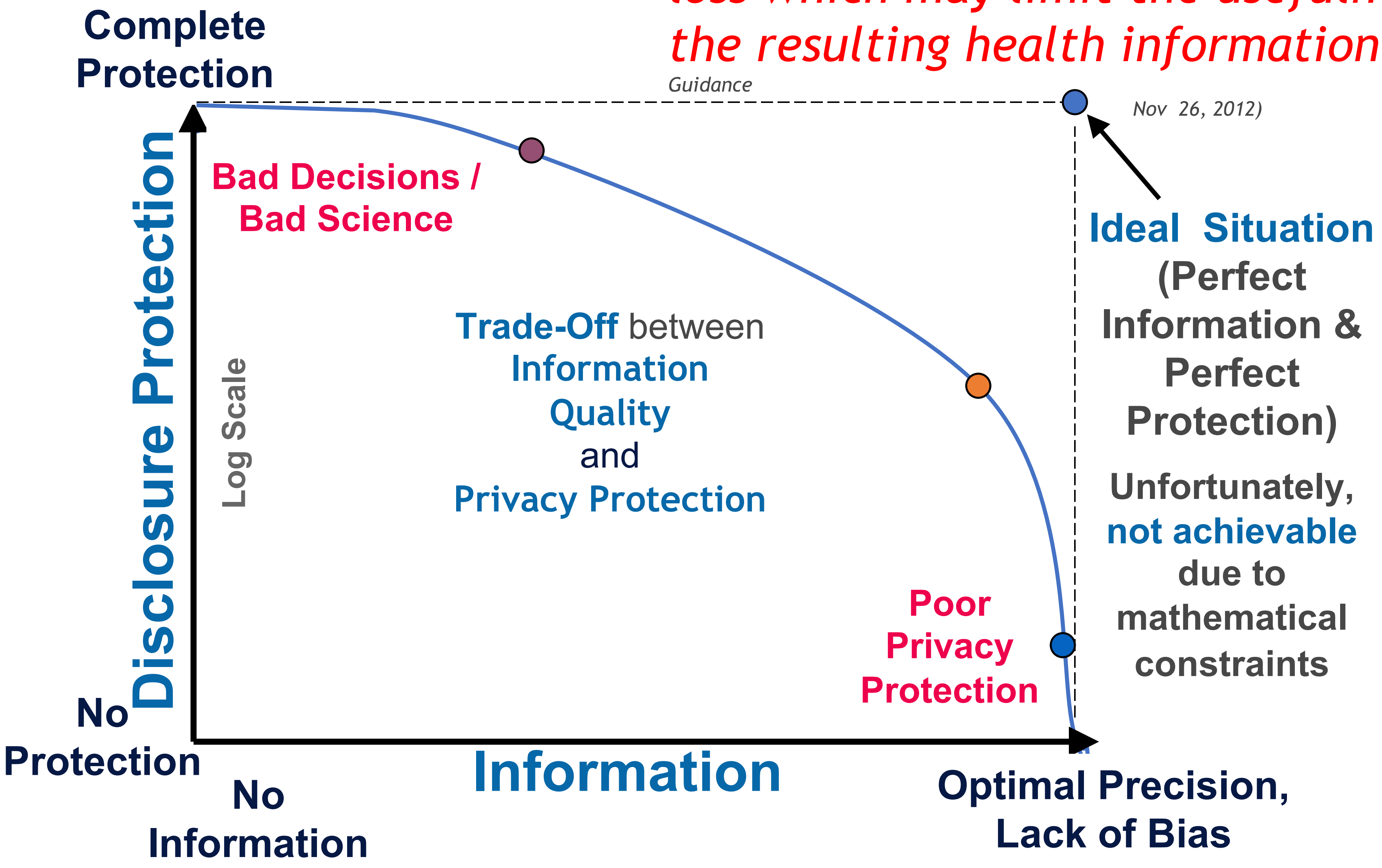
Health Information is not individually identifiable if:

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

(i) Applying such principles and methods, determines that the *risk is very small* that *the information could be used*, alone or *in combination with other reasonably available information, by an anticipated recipient to identify an individual* who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination;

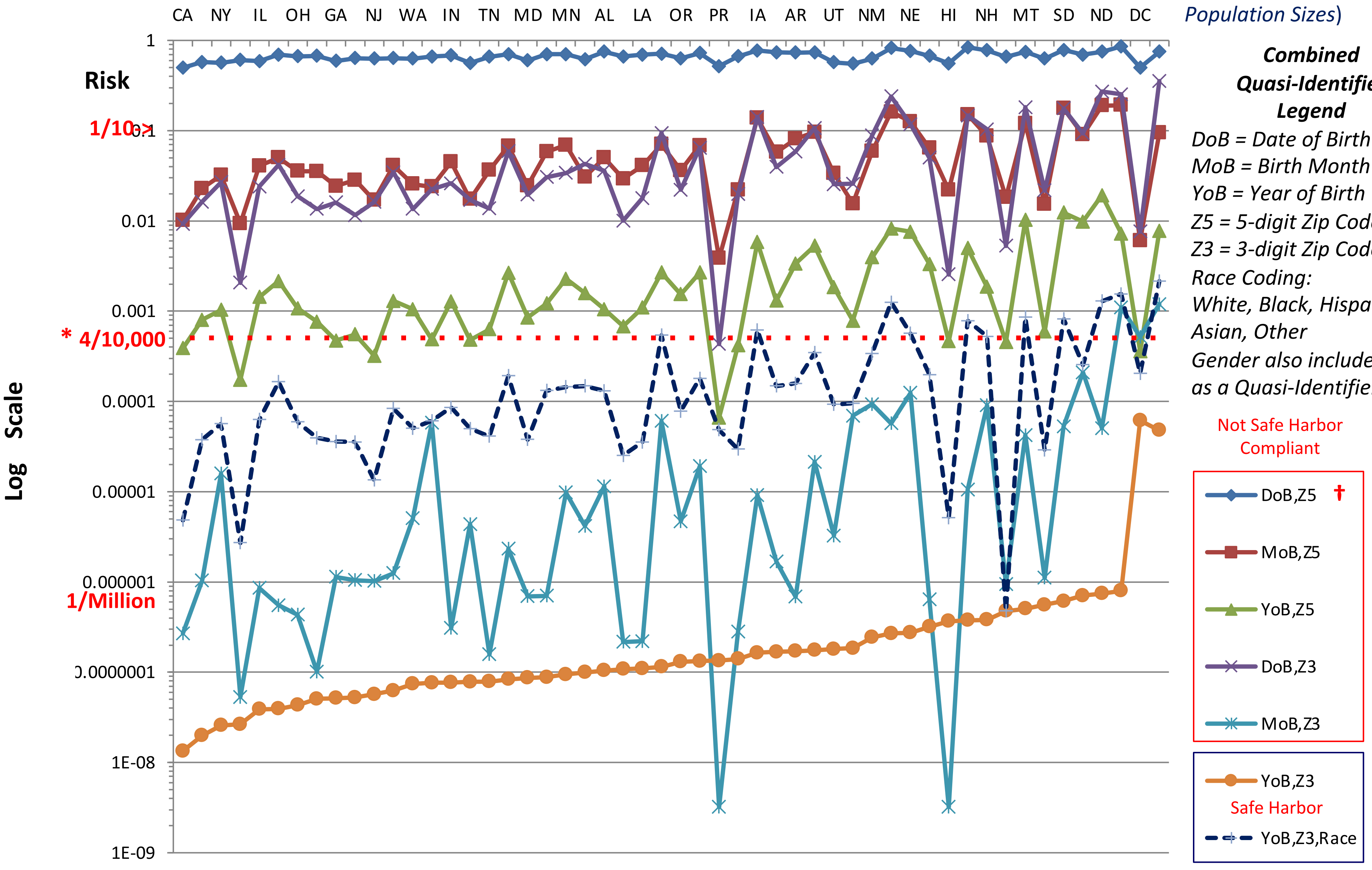
The Inconvenient Truth:

"De-identification leads to information loss which may limit the usefulness of the resulting health information" *(p.8, HHS De-ID Guidance Nov 26, 2012)*



Complete Protection
Disclosure Protection
Log Scale
No Protection
Bad Decisions / Bad Science
Trade-Off between Information Quality and Privacy Protection
Poor Privacy Protection
Information
No Information
Optimal Precision, Lack of Bias
Ideal Situation (Perfect Information & Perfect Protection)
Unfortunately, not achievable due to mathematical constraints

U.S. State Specific Re-identification Risks: Population Uniqueness



(States ordered by Population Sizes)

Combined Quasi-Identifier Legend

DoB = Date of Birth
MoB = Birth Month & Yr
YoB = Year of Birth
Z5 = 5-digit Zip Code
Z3 = 3-digit Zip Code
Race Coding:
White, Black, Hispanic, Asian, Other
Gender also included as a Quasi-Identifier

† HIPAA Safe Harbor does not permit any Dates more specific than the year, or Geographic Units smaller than 3-digit Zip Codes (Z3).

****HIPAA Safe Harbor Risk Estimate***

Graph © DB-J 2013

Data Source: 2010 U.S. Decennial Census

Balancing Disclosure Risk/Statistical Accuracy

- Balancing disclosure risks and statistical accuracy is essential because some popular de-identification methods (e.g. k-anonymity, noise injection) can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses or data mining (distorting variance-covariance matrices, masking heterogeneous sub-groups which have been collapsed in generalization protections)

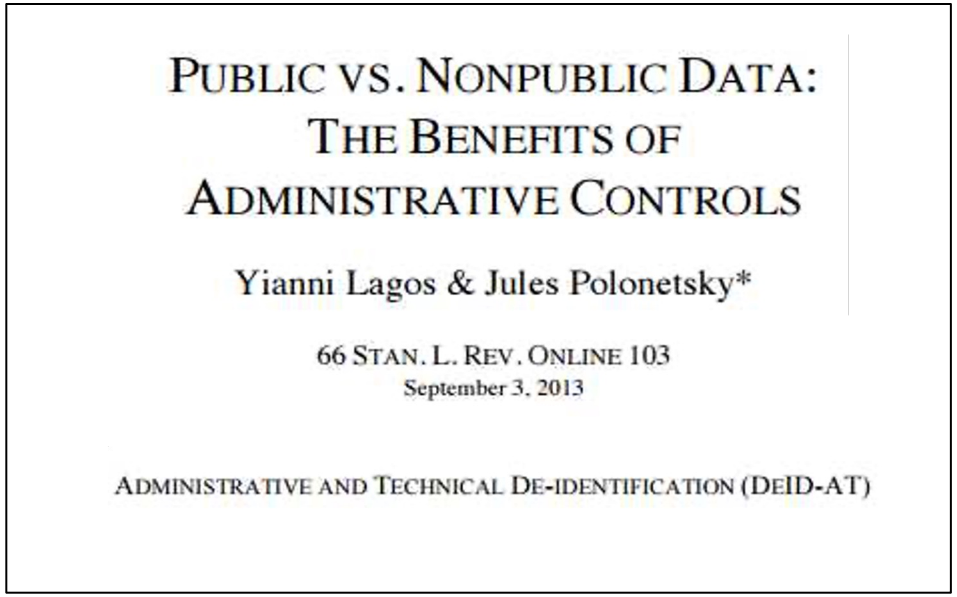
- This problem is well-understood by statisticians, but not as well recognized and integrated within public policy.

- Poorly conducted de-identification can lead to “bad science” and “bad decisions”.

Reference: C. Aggarwal http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf

Supplementing Technical Data De-identification with Legal/Administrative Controls

However, in many cases, because of the possibility of highly-targeted demonstration attacks, arriving at solutions which will appropriately preserve the statistical accuracy and utility will also require that we supplement our statistical disclosure limitation "technical" data de-identification methods with additional legal and administrative controls.



PUBLIC VS. NONPUBLIC DATA: THE BENEFITS OF ADMINISTRATIVE CONTROLS

Yianni Lagos & Jules Polonetsky*

66 STAN. L. REV. ONLINE 103
September 3, 2013

ADMINISTRATIVE AND TECHNICAL DE-IDENTIFICATION (DEID-AT)

Suggested Conditions for De-identified Data Use

Recipients of De-identified Data should be required to:

1) Not re-identify, or attempt to re-identify, or allow to be re-identified, any patients or individuals who are the subject of Protected Health Information within the data, or their relatives, family or household members.

2)Not link any other data elements to the data without obtaining a determination that the data remains de-identified.

3) Implement and maintain appropriate data security and privacy policies, procedures and associated physical, technical and administrative safeguards to assure that it is accessed only by authorized personnel and will remain de-identified.

4) Assure (via internal policies and procedures and contractual commitments for third parties) that all personnel or parties with access to the data agree to abide by all of the foregoing conditions.

And, of course, destructively delete or encrypt the data when no longer needed or in use.

Recommended Skills for De-Identification Expert Teams

- Statistical Disclosure Limitation/Control Theory & Practices
- Privacy Preserving Data Publishing and Mining
- HIPAA/HITECH and Data Privacy Law
- Corporate Compliance and Data Governance
- Medical Informatics and Medical Coding/Billing Systems
- Biostatistics/Epidemiology
- Geographic Information Systems
- Machine Learning/Artificial Intelligence
- Health Systems/Health Economics Research
- Cryptography
- Computer Security
- Data Privacy Computer Science (e.g., Differential Privacy, Homomorphic Encryption)
- Data Management/Architecture Theory and Practices

We also need…

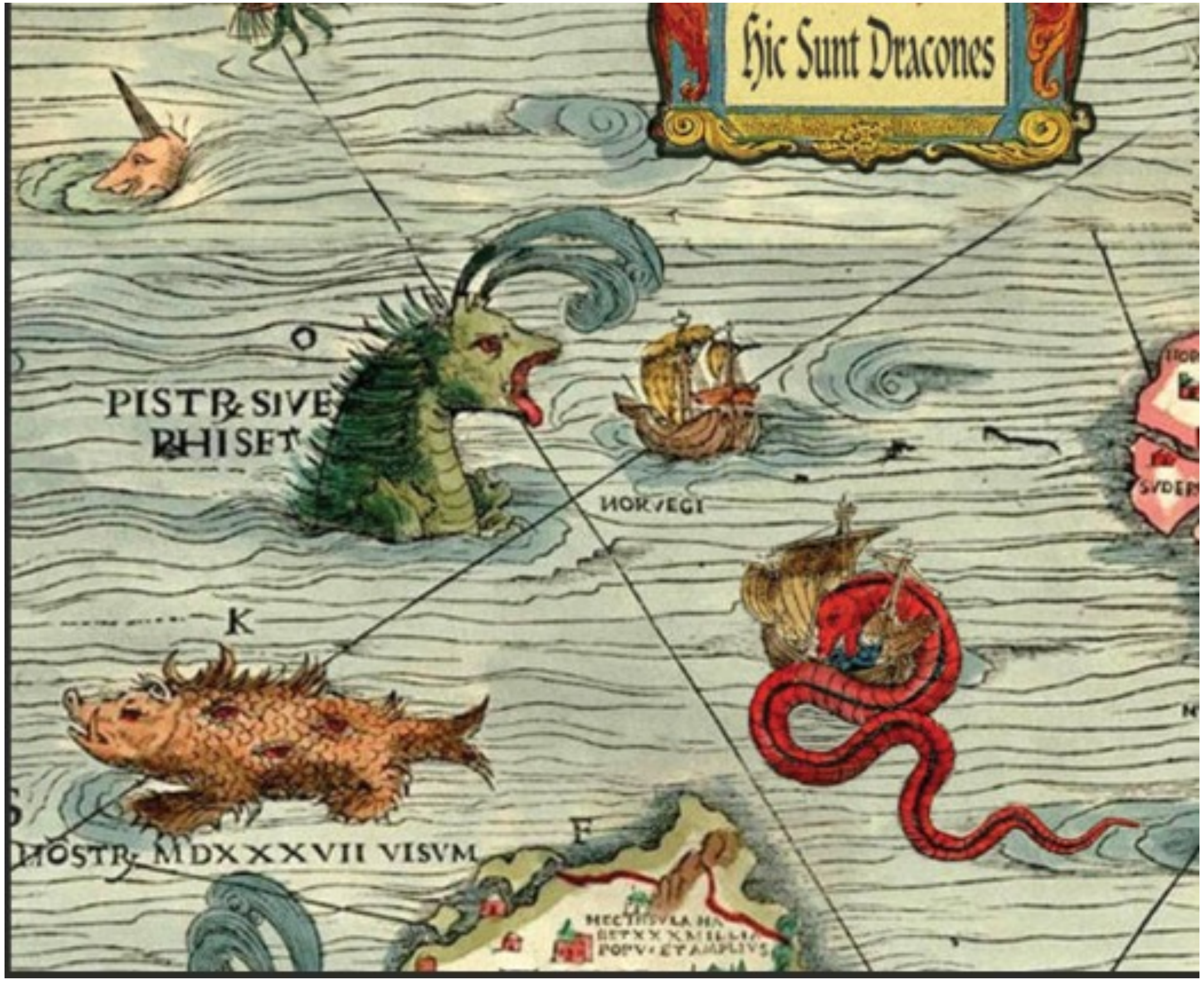
Comprehensive, Multi-sector Statutory Prohibitions Against Data Re-identification

See the new ban on re-identification of de-identified health data under CA AB 718 (2020) –

*Which bans re-identification of previously de-identified health data, **except** where such re-identification is **needed for HIPAA-governed activities**, is required by law, or **where necessary for testing, analysis, or validation of de-identification techniques.***

Should it be applied nationally?

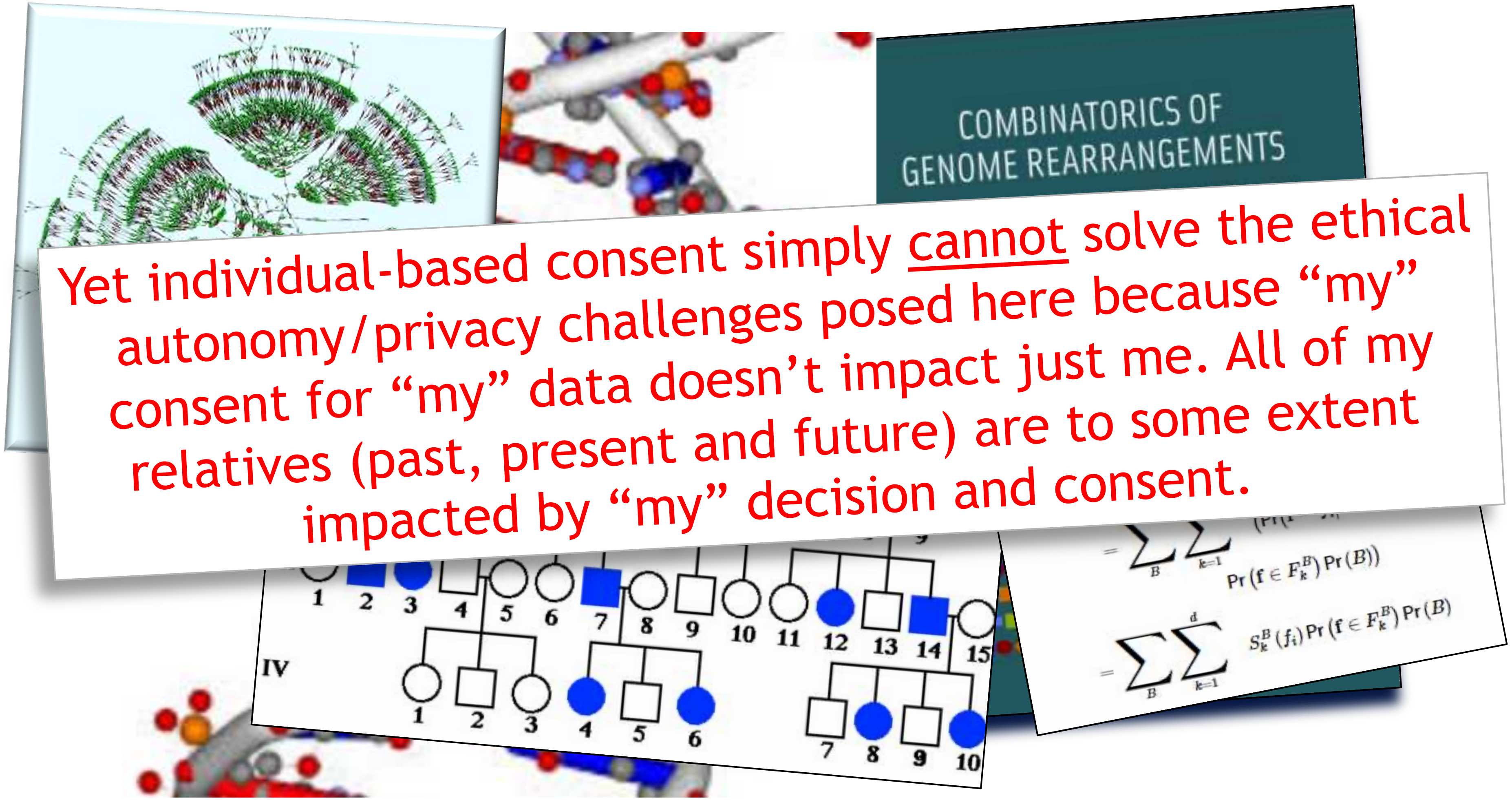
Terra Incognito: *Hic Svnt Dracones*



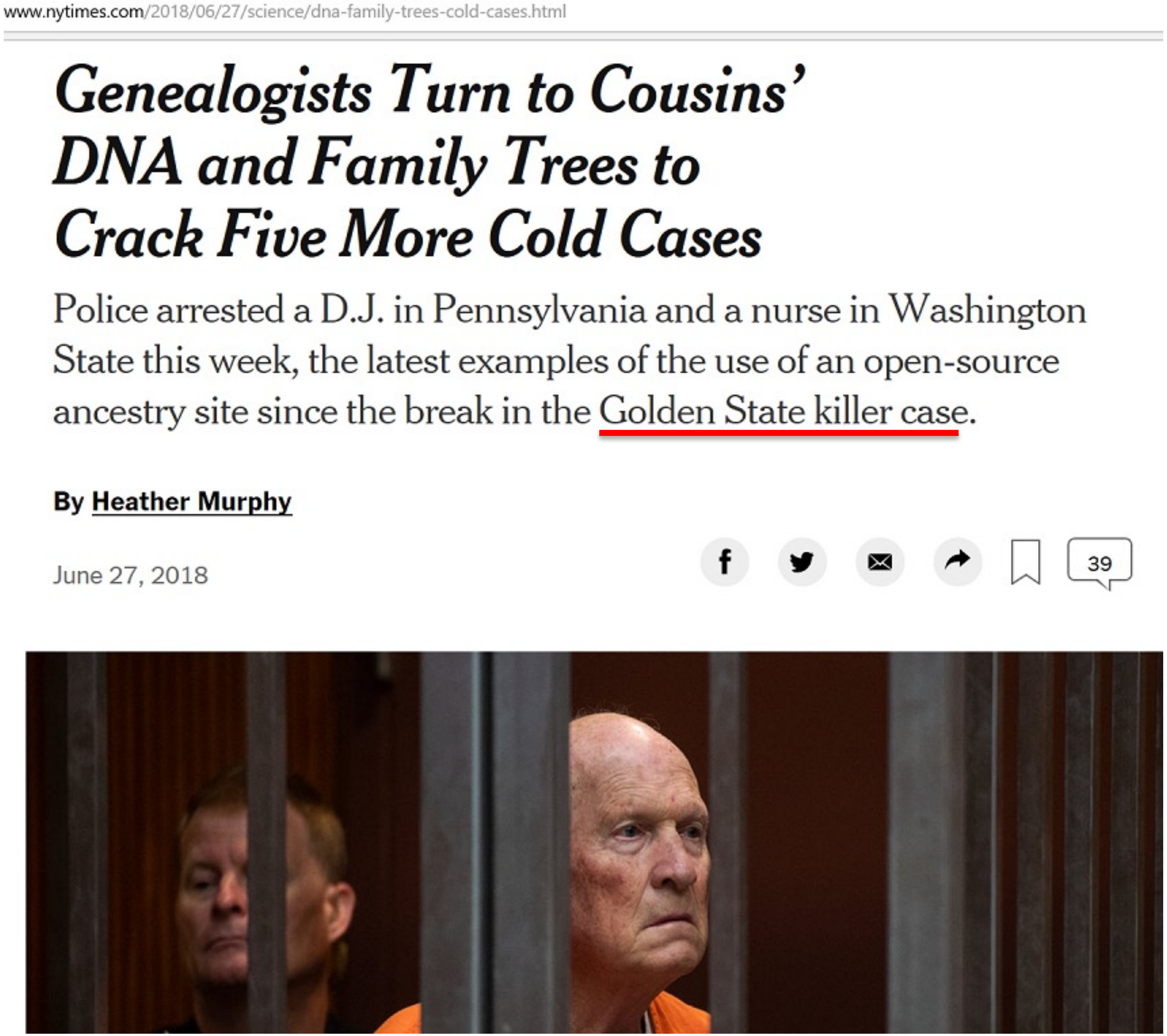
Re-identification Attack News Raises Public Concerns

- Concerns about health privacy for research uses of "de-identified" health and genomic data have repeatedly made national headlines recently, reporting on several highly-publicized attacks and data breaches.
- Importance of the complex ethical and public policy considerations surrounding de-identified data is also accelerating rapidly due to inexpensive whole genome sequencing and widespread advances in artificial intelligence

Given the inherent extremely large combinatorics of genomic data nested within inheritance networks which determine how genomic traits (and surnames) are shared with our ancestors/descendants, the degree to which such information could be meaningfully "de-identified" is non-trivial.



Yet individual-based consent simply cannot solve the ethical autonomy/privacy challenges posed here because "my" consent for "my" data doesn't impact just me. All of my relatives (past, present and future) are to some extent impacted by "my" decision and consent.



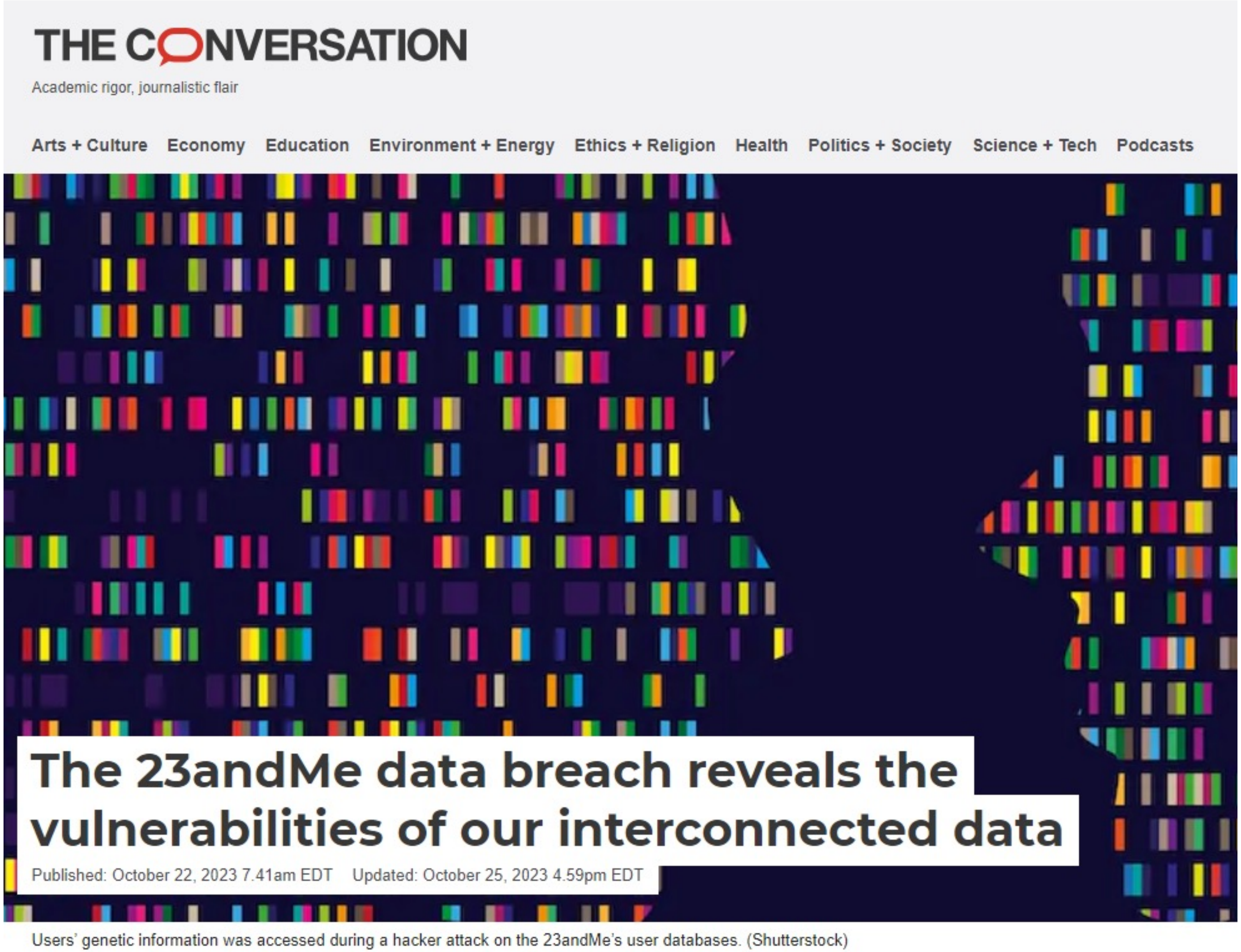
www.nytimes.com/2018/06/27/science/dna-family-trees-cold-cases.html

Genealogists Turn to Cousins' DNA and Family Trees to Crack Five More Cold Cases

Police arrested a D.J. in Pennsylvania and a nurse in Washington State this week, the latest examples of the use of an open-source ancestry site since the break in the Golden State killer case.

By Heather Murphy

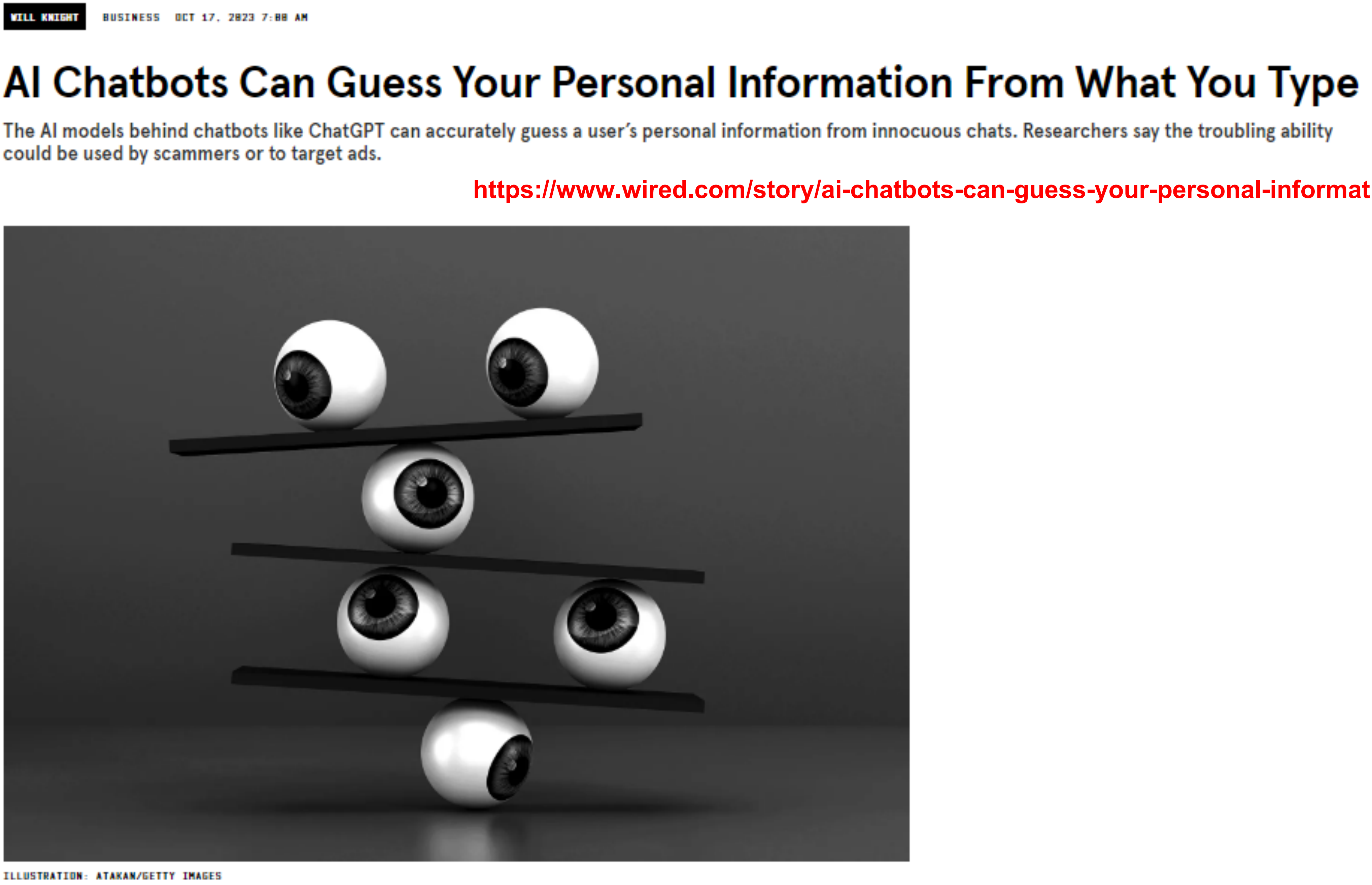
June 27, 2018



The 23andMe data breach reveals the vulnerabilities of our interconnected data

Published: October 22, 2023 7.41am EDT Updated: October 25, 2023 4.59pm EDT

Users' genetic information was accessed during a hacker attack on the 23andMe's user databases. (Shutterstock)



WILL KNIGHT BUSINESS OCT 17, 2023 7:00 AM

AI Chatbots Can Guess Your Personal Information From What You Type

The AI models behind chatbots like ChatGPT can accurately guess a user's personal information from innocuous chats. Researchers say the troubling ability could be used by scammers or to target ads.

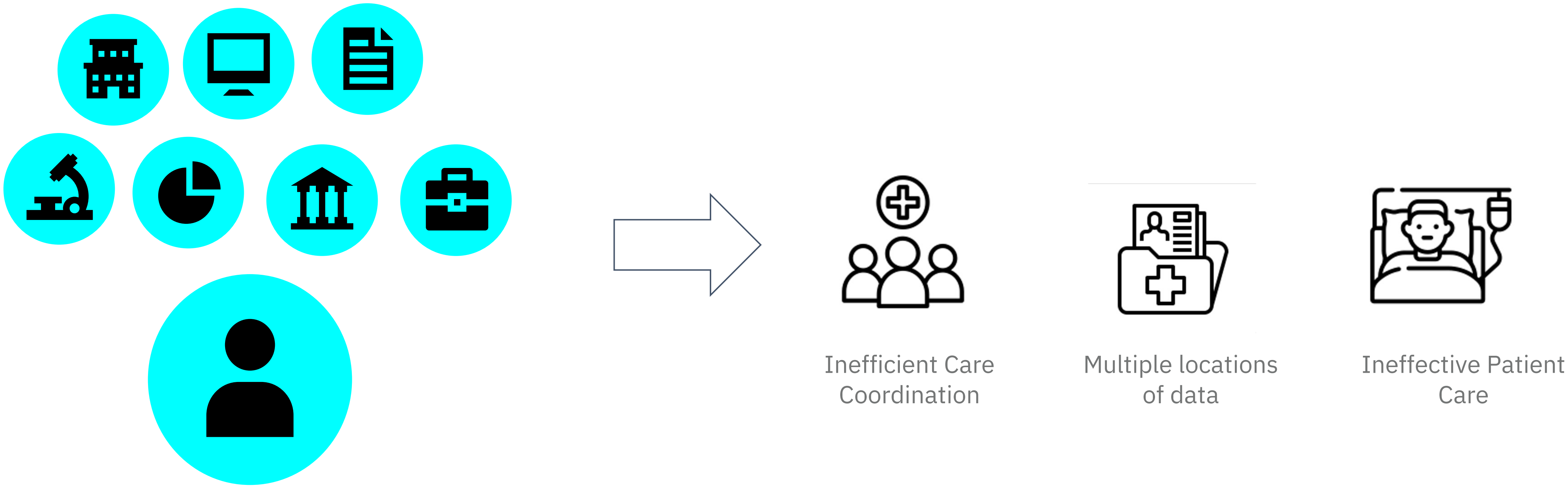
ILLUSTRATION: ATAKAN/GETTY IMAGES

https://www.wired.com/story/ai-chatbots-can-guess-your-personal-information/

How are Researchers Working with Health Data Today?

Claire Manneh

State of Health Data & Health-related Data



Patient data is fragmented and a patient’s health journey is not connected

Inefficient Care Coordination

Multiple locations of data

Ineffective Patient Care

Historically, connecting health data was manual and time-intensive

1. Find data partners by word-of-mouth
2. Get counts of patients of interest from every possible partner
3. Send detailed cohort criteria (ICD codes, histology, pathology, etc.)
4. Partner runs SAS queries and sends back report
5. Sign BAA with partner
6. Partner sends data to you
7. Prepare cuts of your data for comparisons
8. Create homegrown tokenization (salt / hash / encryption) to compare overlap or hire consultant
9. Work with independent expert on HIPAA risk disclosure assessment
10. Continue refreshing data

Connecting data improves population health outcomes

NEED

A large health system wants to develop a population health initiative to address socioeconomic barriers to care and the impact on outcomes using their internal EHR data.

Hospital/ Health System

FIRST PARTY DATA

- Electronic Health Records

THIRD PARTY DATA

- Social Determinants Data
- Insurance Claims Data

Questions

- What socioeconomic dynamics impact patient access and outcomes?
- How do those dynamics vary by site/locations across the health system?

SOLUTION

Using the Datavant Switchboard, the health system can de-identify and connect their electronic health records to third-party data, and develop interventions based on population insights.

CONNECTED DATA

- Electronic Health Records
- Social Determinants Data — *Assess impact of income and employment on access to care*
- Insurance Claims Data — *Understand medical history and resource utilization*

Based on trends related to income, transportation access and food insecurity by county, the health system can deploy interventions such as partnering with local food banks and ride-sharing apps to improve patient access to care and health outcomes.

Insights on Patient 956

- Recently unemployed, doesn't own a car, lives in a food desert
- Diagnosed last month with Type II Diabetes
- 2 ER admissions in last 12 months

Record Linkage

HIPAA

- Prohibits the sharing of identifiable individual health information outside of established legal pathways (TPO, public health, etc)

PHI/PII

- Without identifying information, it's difficult or impossible to link patient records – within a data set, and more so across data sets, let alone across data suppliers

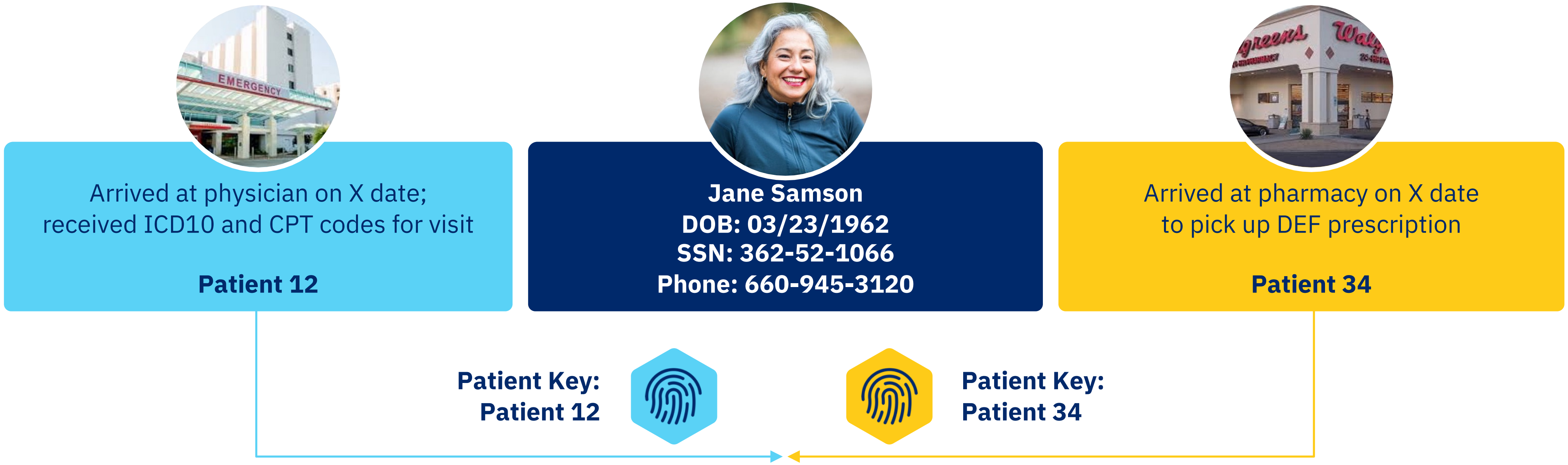
Purpose

- Need in advanced data uses for researchers to link data from different sources about the same patient, even though there's no need to know who the patient is

Tokenization: A potential solution

Privacy-Preserving Record Linkage (PPRL)

Cryptographic method of representing identity in a de-identified manner while preserving ability to link health data



Arrived at physician on X date;
received ICD10 and CPT codes for visit

Patient 12

Jane Samson
DOB: 03/23/1962
SSN: 362-52-1066
Phone: 660-945-3120

Arrived at pharmacy on X date
to pick up DEF prescription

Patient 34

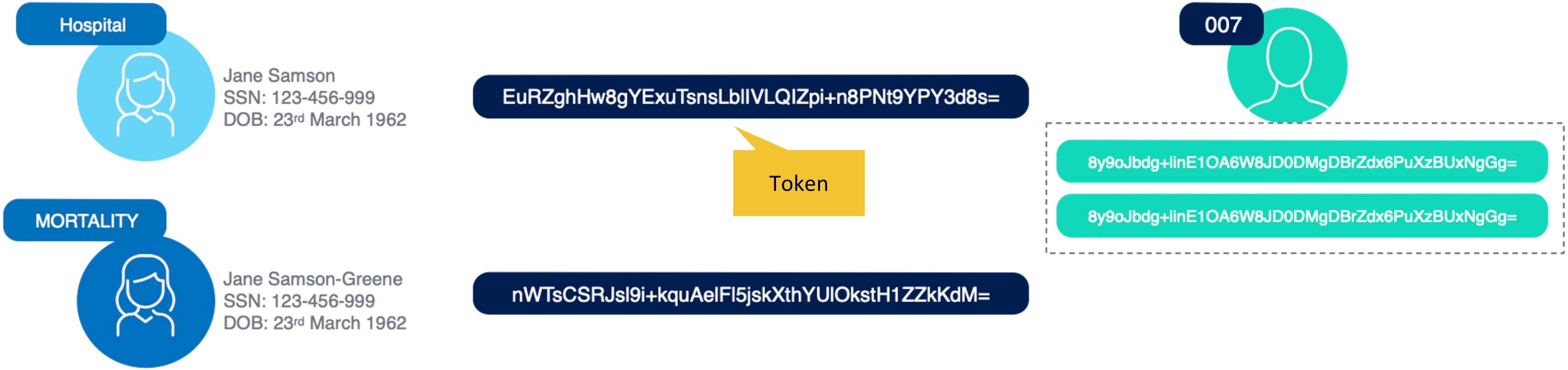
Patient Key:
Patient 12

Patient Key:
Patient 34

PPRL matches Patient 12 and Patient 34
as the same patient to get the complete picture

What Does PPRL Look Like in Practice?

Identifiable demographic attributes about an individual are extracted by the data holder → **De-identified "tokens" are generated** using privacy-preserving software → **Authorized recipient links and matches** de-identified tokens across multiple data partners



Hospital

Jane Samson
SSN: 123-456-999
DOB: 23rd March 1962

EuRZghHw8gYExuTsnsLblIVLQIZpi+n8PNt9YPY3d8s=

Token

MORTALITY

Jane Samson-Greene
SSN: 123-456-999
DOB: 23rd March 1962

nWTsCSRJsl9i+kquAelFl5jskXthYUlOkstH1ZZkKdM=

007

8y9oJbdg+linE1OA6W8JD0DMgDBrZdx6PuXzBUxNgGg=

8y9oJbdg+linE1OA6W8JD0DMgDBrZdx6PuXzBUxNgGg=

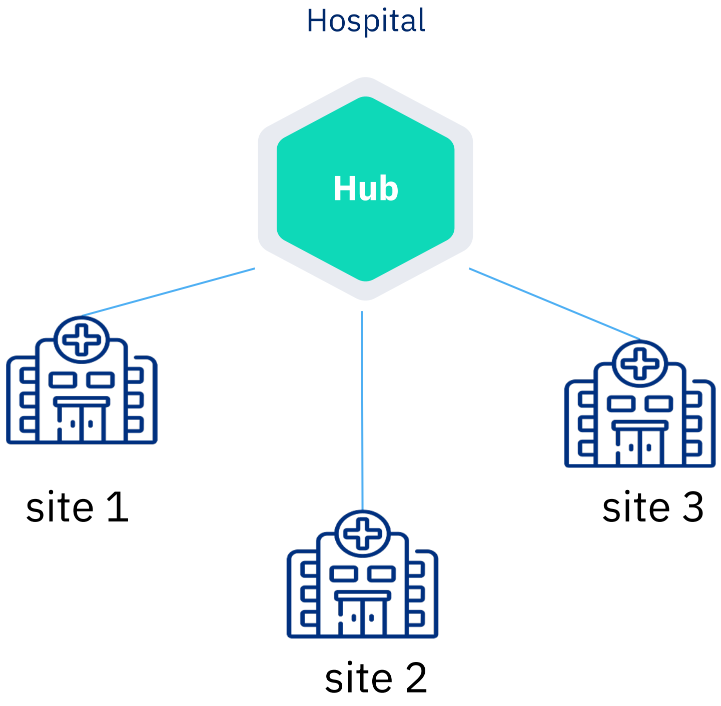
The Juicer Analogy

Green Apple = First Name
Red Apple = Last Name
Carrot = Date of Birth
Turmeric = Street Address
Beet = Zip Code

Juice = Token

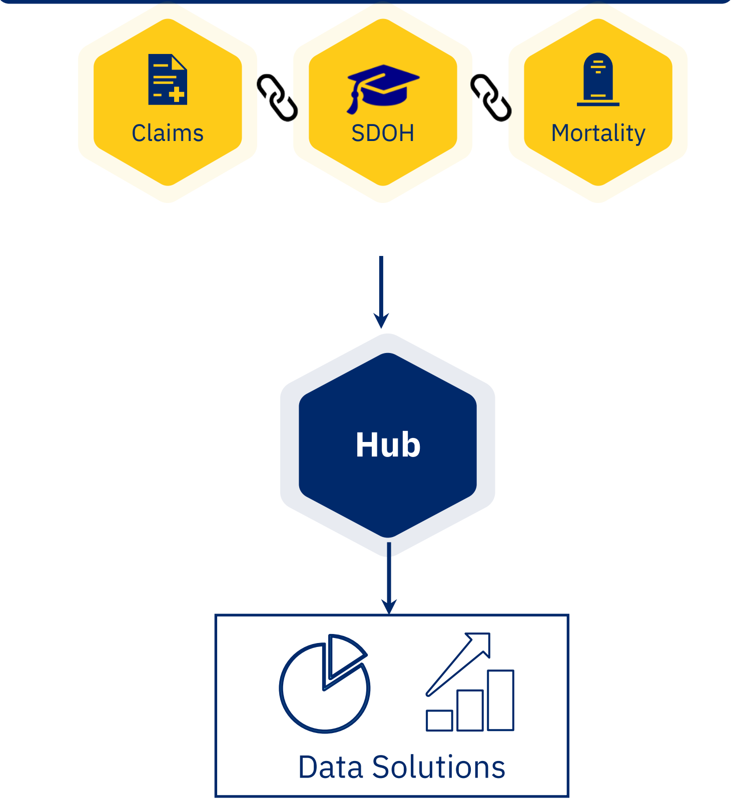
Tokenization allows linking of a patient’s records to build a longitudinal view of their journey

Connect and Add Sites



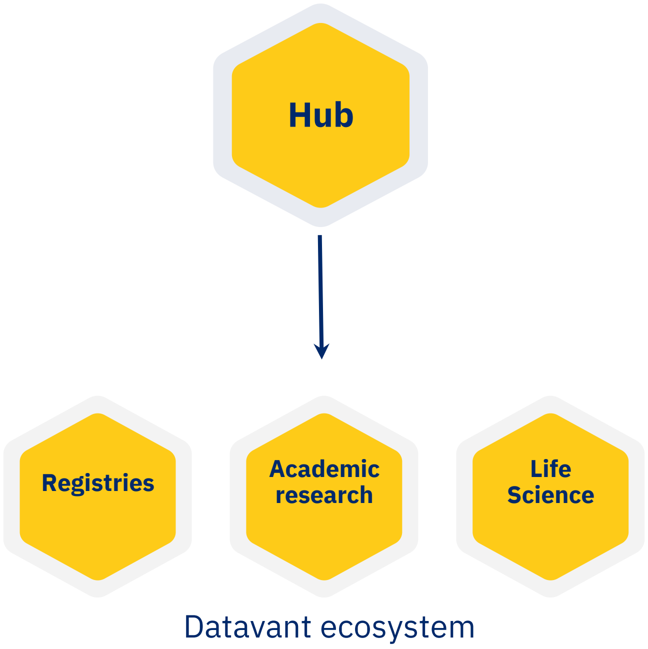
De-identification using **privacy-preserving linkages** unlocks possibilities for enrichment and partnership

Data Enrichment & Discovery



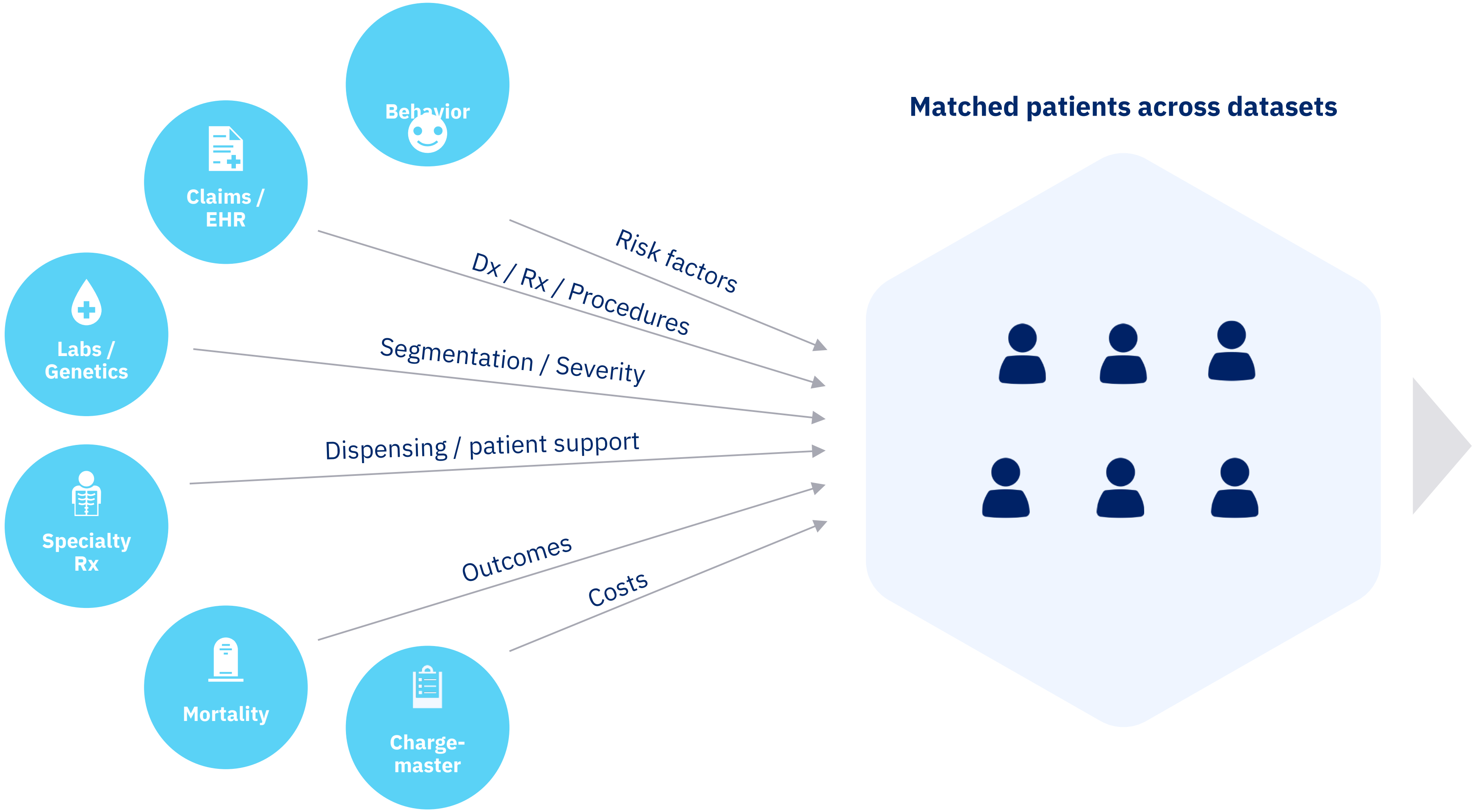
Discover relevant data partners or cohorts, **enrich data** with ecosystem partners

License Data to Others



Share privacy-preserved, de-identified data available **on your terms**

The comprehensive and more complete datasets are suitable to answer a variety of questions…



- Epidemiology / segmentation
- Rare disease identification
- Trial feasibility
- Patient / Provider targeting
- Brand tracking
- HEOR

What’s happening now with tokenization and linking in the real world?

Linkage and Unification Cross-Repository

National Institutes on Health has multiple repositories with different data types about the same population

NIH National Center for Advancing Translational Sciences

N3C

National COVID Cohort Collaborative

(largest collection of secure and deidentified clinical data in the United States for COVID-19 research)

NIH

All of Us RESEARCH PROGRAM

Collection of study data from 1m+ people in the US

PCORnet, National Patient-Centered Clinical Research Network

Encrypted tokenization across these networks allow over 60 hospitals to link their EHR data in a privacy preserving way

pcornet
The National Patient-Centered Clinical Research Network
Columbia Vagelos College of Physicians and Surgeons
Weill Cornell Medicine
Einstein
Albert Einstein College of Medicine
Montefiore Doing More
Mount Sinai
NYU Langone Health
Emory University
DukeHealth
The University of North Carolina at Chapel Hill
Vanderbilt Health Affiliated Network
Meharry Medical College
Vanderbilt University Medical Center
UF University of Florida
Health Sciences South Carolina
A Collaborative To Advance Health Sciences
Mayo Clinic
Wake Forest Baptist Health

80 million+ individuals

Longitudinal data 2009-2023

8 clinical networks, 2 health plans

70 health systems

337 hospitals

1,024 community clinics

3,564 primary care practices

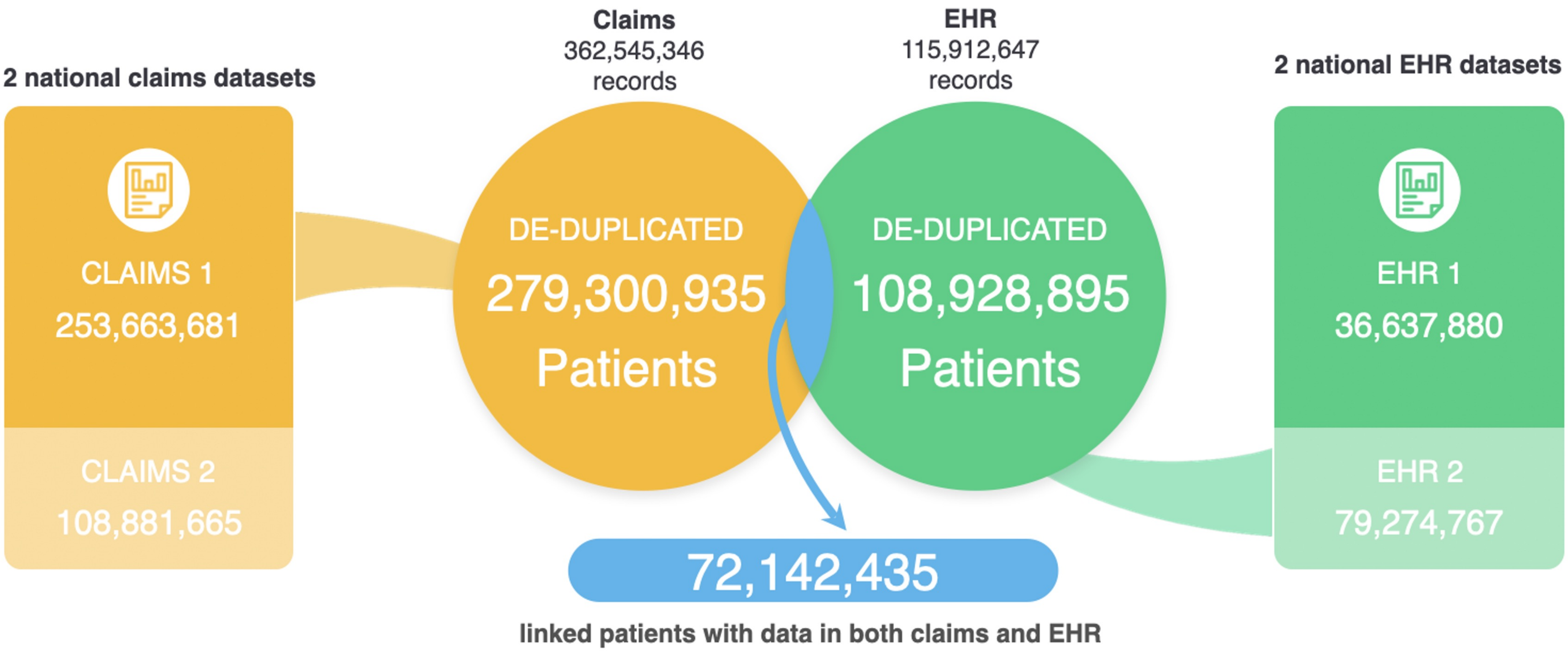
338 emergency departments

COVID-19
RESEARCH DATABASE
https://covid19researchdatabase.org/

A centralized repository of de-id tokenized datasets encompassing EHR, long-term care, claims, SDoH

REAGAN-UDALL FOUNDATION for the Food and Drug Administration

2021 Innovations in Regulatory Science Awards



2 national claims datasets

CLAIMS 1
253,663,681

CLAIMS 2
108,881,665

Claims
362,545,346 records

DE-DUPLICATED
279,300,935
Patients

EHR
115,912,647 records

DE-DUPLICATED
108,928,895
Patients

2 national EHR datasets

EHR 1
36,637,880

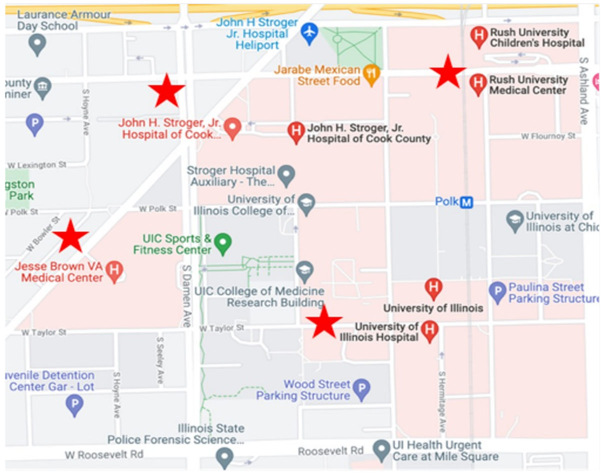
EHR 2
79,274,767

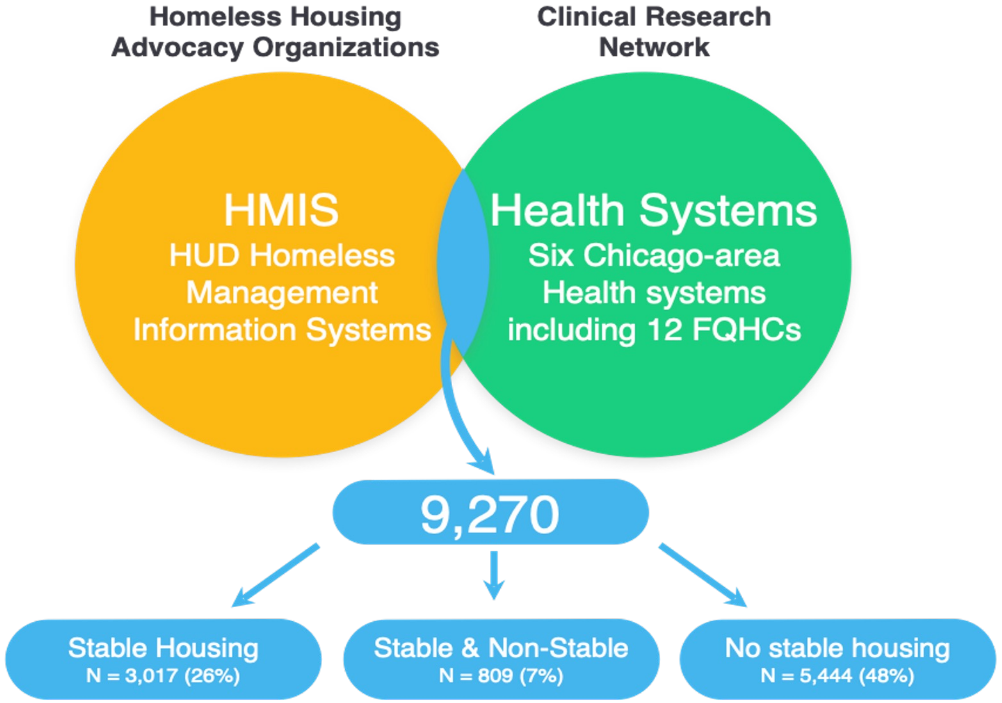
72,142,435

linked patients with data in both claims and EHR

Novel Linkages for Vulnerable Populations

Novel linkages for vulnerable populations


Laurance Armour Day School
John H Stroger Jr. Hospital Heliport
Rush University Children's Hospital
Jarabe Mexican Street Food
Rush University Medical Center
John H. Stroger, Jr. Hospital of Cook...
John H. Stroger, Jr. Hospital of Cook County
W Flournoy St
W Lexington St
Stroger Hospital Auxiliary - The...
University of Illinois College of...
Polk
University of Illinois at Chic
W Polk St
UIC Sports & Fitness Center
Jesse Brown VA Medical Center
UIC College of Medicine Research Building
University of Illinois
Paulina Street Parking Structure
University of Illinois Hospital
W Taylor St
S Damen Ave
S Ashland Ave
Wood Street Parking Structure
Juvenile Detention Center Gar - Lot
Illinois State Police Forensic Science...
W Roosevelt Rd
Roosevelt Rd
UI Health Urgent Care at Mile Square


Homeless Housing Advocacy Organizations

Clinical Research Network

HMIS
HUD Homeless Management Information Systems

Health Systems
Six Chicago-area Health systems including 12 FQHCs

9,270

Stable Housing
N = 3,017 (26%)

Stable & Non-Stable
N = 809 (7%)

No stable housing
N = 5,444 (48%)

Data linkages were used to surface homeless individuals who sought care within health systems, including frequently health systems that were close to another. Area health systems formulated a **homeless housing subsidy program** to help get individuals back into housing with the support of their advocacy organizations. Linked data were mined to understand comorbidities.

Source:
Joining Healthcare and Homeless Data Systems using Privacy-Preserving Record Linkage Software. https://ajph.aphapublications.org/doi/10.2105/AJPH.2021.306304
Variability in comorbidities and health services use across homeless typologies: multicenter data linkage between healthcare and homeless systems. https://doi.org/10.1186/s12889-021-10958-8

Break

De-Identification

and

The Law(s)

De-Identification Under the New State Laws

Ann Waldo

STANDARDS

- **Play a vital role globally by facilitating communication, innovation, progress**
- Early civilizations developed standardized ways to measure time and space – calendars, clocks, units of length, weight, etc. Some idiosyncratic (*e*.g., King of England's own arm became the standard in 1120 AD)
- **Int'l trade and Industrial Revolution made greater standardization essential**

Consider calendars – Roman, Mayan, Egyptian, Islamic, Hebrew, Hindu, Persian…

→→→→ Gregorian calendar introduced in 1582, widely adopted by 19th century, now **the** international civil standard used worldwide

But what about de-identification standards? State laws are taking us backward to the realm of inconsistent standards

CA CCPA (Original)

- Original CCPA had a novel definition of “deidentification” that applied to ALL data – and wasn’t at all harmonized with HIPAA standard
- No exception for HIPAA de-ID’d data
- While meeting both the HIPAA and the CCPA de-ID’n standards would have been possible, it was also possible to not meet both. Would have resulted in painful and expensive lawyering, contractual wrangling over risk, delays, costs, litigation risk, etc.

- Two-year effort to change CA law to harmonize de-ID’n with HIPAA for patient information
  - *Successful!*
  - *Multi-stakeholder collaboration, including privacy advocates*
  - *CA AB 713 (2020)*

De-ID’n under CA Law Today*

- ***De-ID’n for patient information in CA now harmonized with HIPAA de-ID’n**
  - “Patient information” is broadly defined (“PHI Plus”)
  - Does include medical data, does not include consumer health data (smart watches, etc.)

- **NOTE - All data that is not patient information is subject to the general CCPA definition**, not harmonized with HIPAA.

- **Some new provisions apply to de-ID’d patient information**

Okay, that’s CA.

What about the other new state consumer privacy laws??

14 of the 15 enacted to date (i.e., all except Delaware)
have a two-tier structure similar to CA’s:

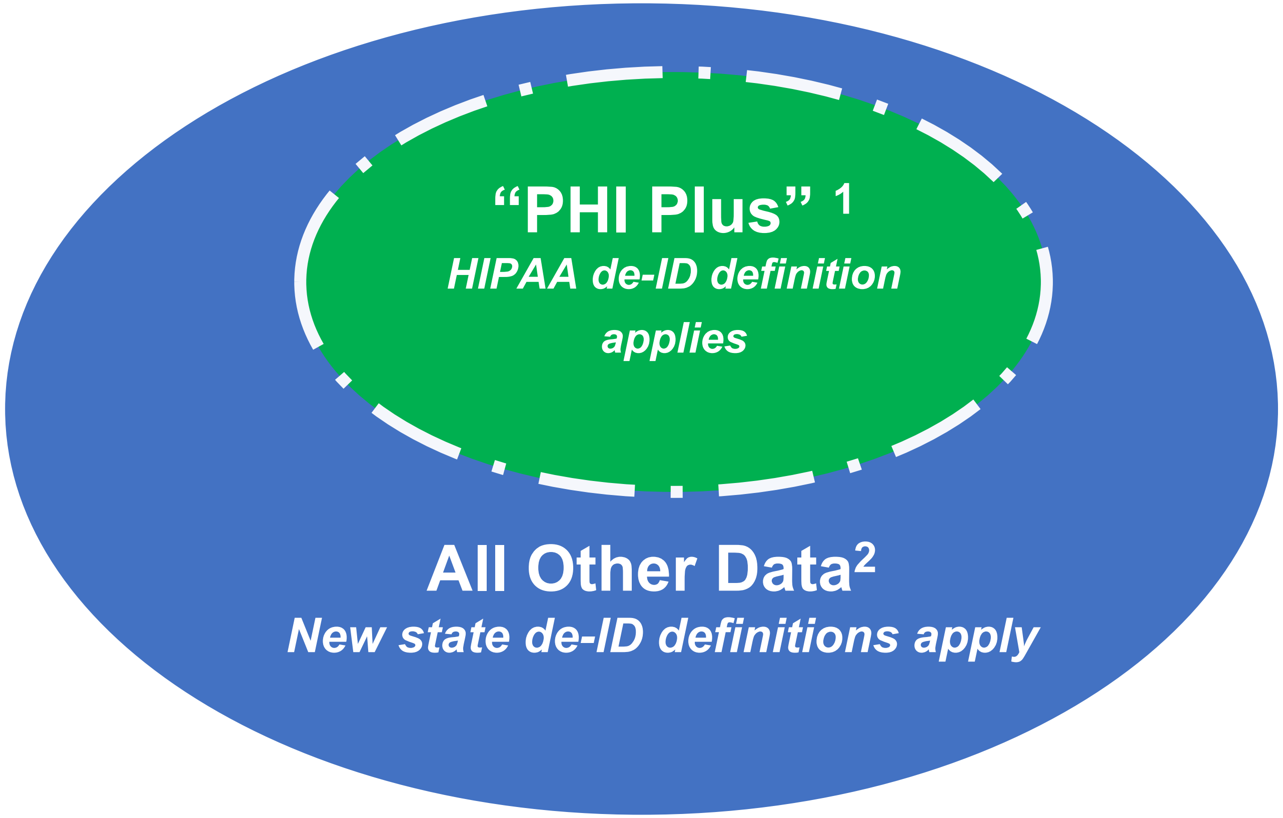
- **HIPAA de-ID’n applies to “PHI Plus” (PHI plus other medical data)**
- **New state-specific de-ID definition applies to all other data**

*Treating WA’s and NV’s new “consumer health” laws as general privacy laws here due to their breadth of scope

Which state De-ID standard applies to which data? For 14 of the 15 state laws…

[1]**"PHI Plus"** is "patient information" in CA law and has other designations under 13 other state laws. Refers to PHI plus other specified medical data. Examples include PHI, research data subject to Common Rule, Part 2 data, etc. Note – the exact perimeters of what's included in "PHI Plus" data vary by state (hence the jagged line here.)

[2]"**All Other Data**" refers to all data not included in the exemption for "PHI Plus" data. Examples include consumer health data, SDOH, demographic data, etc.


"PHI Plus" 1
HIPAA de-ID definition applies
All Other Data2
New state de-ID definitions apply

More *complexities with de-ID’n under the 14 new state laws (excluding Delaware)*

- The perimeter of the inner circle – the “PHI Plus” subject to HIPAA de-ID’n – varies by state
- The de-ID’n language applicable to data in the outer circle varies by state
- Some of the actual definitions include business conduct requirements; some do not

Example of harmonized de-identification standard (CA)

[Exempt data includes]

(A) Information that meets both of the following conditions:

(i) It is deidentified in accordance with the requirements for deidentification set forth in Section 164.514 of Part 164 of Title 45 of the Code of Federal Regulations.

(ii) It is derived from patient information that was originally collected, created, transmitted, or maintained by an entity regulated by the Health Insurance Portability and Accountability Act, the Confidentiality Of Medical Information Act, or the Federal Policy for the Protection of Human Subjects, also known as the Common Rule.

Example of a new general de-identification definition (CO)

"De-identified data" means data that cannot reasonably be used to infer information about, or otherwise be linked to, an identified or identifiable individual, or a device linked to such an individual, if the controller that possesses the data:

(a) Takes reasonable measures to ensure that the data cannot be associated with an individual;

(b) Publicly commits to maintain and use the data only in a De-identified fashion and not attempt to re-identify the data; and

(c) Contractually obligates any recipients of the information to comply with the requirements of this subsection (11).

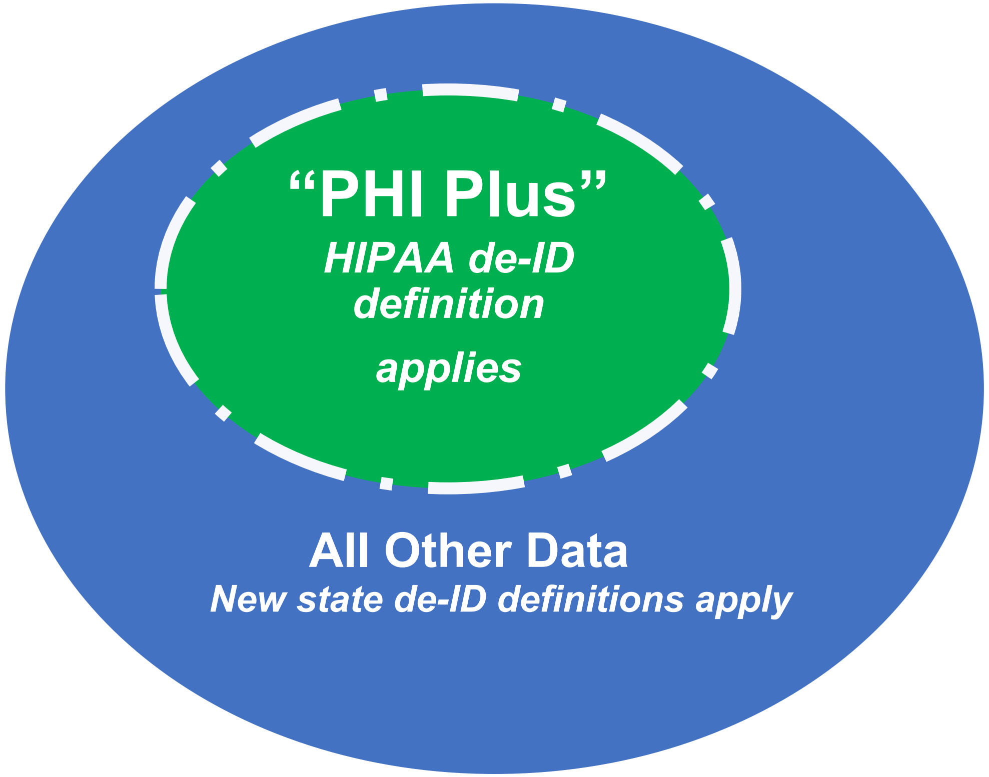
But wait….

What about Delaware??

Delaware's privacy law:

- **Is the ONLY state that does not recognize the HIPAA de-ID standard - not even for PHI**
- Does NOT have a two-tier de-ID structure similar to CA's
- Delaware's general de-ID definition applies to ALL data. It's like the original CCPA (modified in 2020 to harmonize de-ID with HIPAA for "PHI Plus")

14 of 15 State Privacy Laws



"PHI Plus"

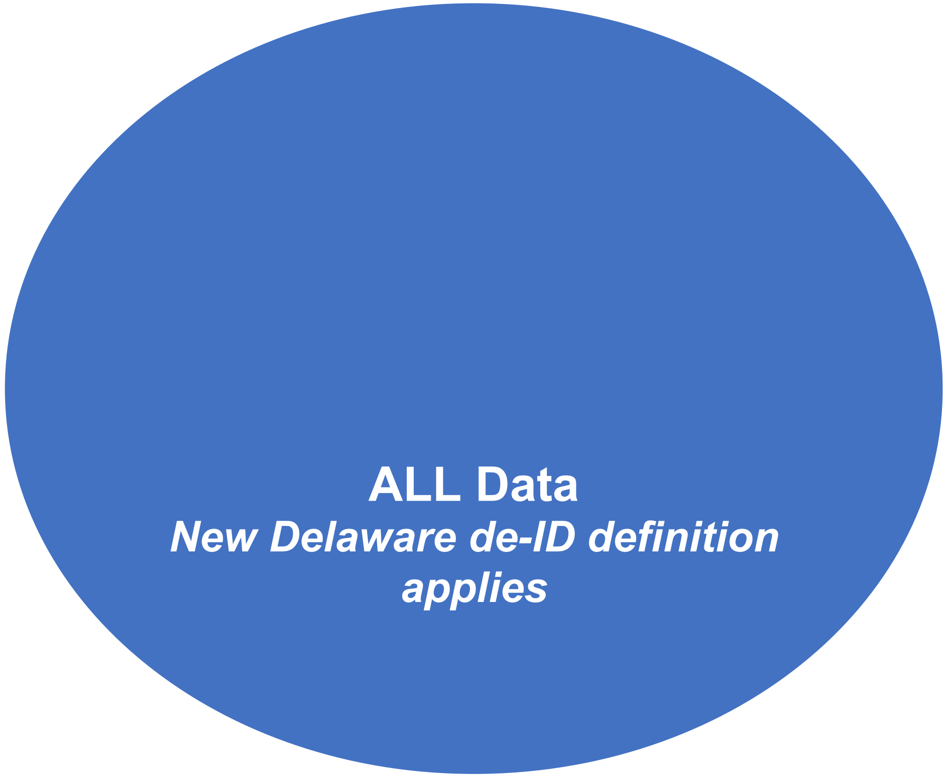
HIPAA de-ID definition

applies

All Other Data

New state de-ID definitions apply

Delaware Privacy Law



ALL Data

New Delaware de-ID definition applies

Audience Questions

- *How do you think that compliance with all the varying U.S. de-ID'n standards can be achieved? How can such be substantiated?*
- *What are the likely ramifications of Delaware not recognizing HIPAA de-ID?*

Other New State Law Provisions Regarding De-ID'n

1) **CA Ban on re-identification of de-ID'd patient information**

- Cannot re-identify, or attempt to re-identify, de-ID'd patient information (data exempt from CCPA because of newly harmonized de-ID'd definition)
- Exceptions to the ban:
  - TPO under HIPAA (Treatment, Payment, Operations)
  - Public Health under HIPAA
  - Research done in accordance with HIPAA or Common Rule
  - Under a contract to test or validate de-ID'n, provided other uses are banned
  - If required by law

  Note – no other exceptions, including for "white hat" researchers, journalists, etc.

- **Scope - a business or other person ---i.e., broader than the rest of the law's scope**

Other New State Provisions Regarding De-ID’n

2) CA Contractual Requirements for Sales

- A contract for the sale or license of de-ID’d patient information must include the following (or substantially similar) terms:
  - Statement about inclusion of de-ID’d patient info
  - Ban on re-ID’n and attempted re-ID’n
  - Downstream contractual terms that are same or stricter
- Scope - one of the parties resides or does business in CA

Other New State Provisions Regarding De-ID’n

3) CA Privacy Notice Requirements

- Scope - a business (per CCPA)
- If a business sells or discloses de-ID’d patient information that’s exempt from CCPA because of the newly harmonized de-ID’d definition for health data, then it must include in its Privacy Policy:

  (a) a statement that it sells or discloses de-ID’d patient information, and

  (b) whether it uses one or more of:
  the HIPAA Safe Harbor method, or
  the expert determination method.

Other New State Provisions Regarding De-ID'n

4) CA - Applicable Law Applies to Re-ID'd Data

- Scope - a business (per CCPA)
- Data that was exempt from CCPA because it qualified for the newly harmonized de-ID'd definition for patient information, *but then became re-identified,* becomes subject to applicable privacy law, including HIPAA, CA CMIA, or CCPA, if applicable

Other New State Provisions re: De-ID

5) Pseudonymization makes its first appearance in US law

- Several states now define pseudonymization *a la* GDPR
- If data is properly pseudonymized, certain state obligations don’t apply.
- And some new requirements apply to pseudonymized data
- *Again – the problem is inconsistency – not all new state laws recognize pseudonymization at all*

Other New State Provisions Regarding De-ID’n

6) Multiple States – New Oversight Duties

- Controller that discloses de-ID’d data must:
  - Exercise reasonable oversight to monitor the data recipients’ compliance with contractual commitments re: the data
  - Take appropriate steps to address any breach of the contractual commitments

- *Some states apply these oversight duties only to de-ID’d data; some to both de-ID’d and pseudonymized data*

Other New State Provisions Regarding De-ID'n

7) Multiple States – Benefits of De-ID'd Data

- Some states allow the use of de-ID'd data to be a factor taken into account in Data Protection Assessments
- Some states have this provision for both de-ID'd and pseudonymized data; some just de-ID'd data

Potential Consequences

As Divergent Definitions of De-Identification Are Enacted

- FUD – fear, uncertainty, doubt
- Administrative and legal costs
- Delays, friction, contracting obstacles
- Burdens on medical research, medical progress
- Harm to patients and the public

Help educate policymakers about importance of harmonizing de-ID'n

Share best practices re: compliance with de-ID'n standards

EU Perspectives on Anonymisation and Pseudonymisation (…with a Life Sciences focus)

Andrew Kopelman

Topics

- Three threshold questions on anonymisation:

  Separate legal basis for anonymisation?

  Does anonymisation need to be ’absolute’?

  Does anonymisation depend on the ‘holder’?

- Enter *SRB v. EDPS (2023)* : back to contextuality / relativity
- Implications and choices; pseudonymisation

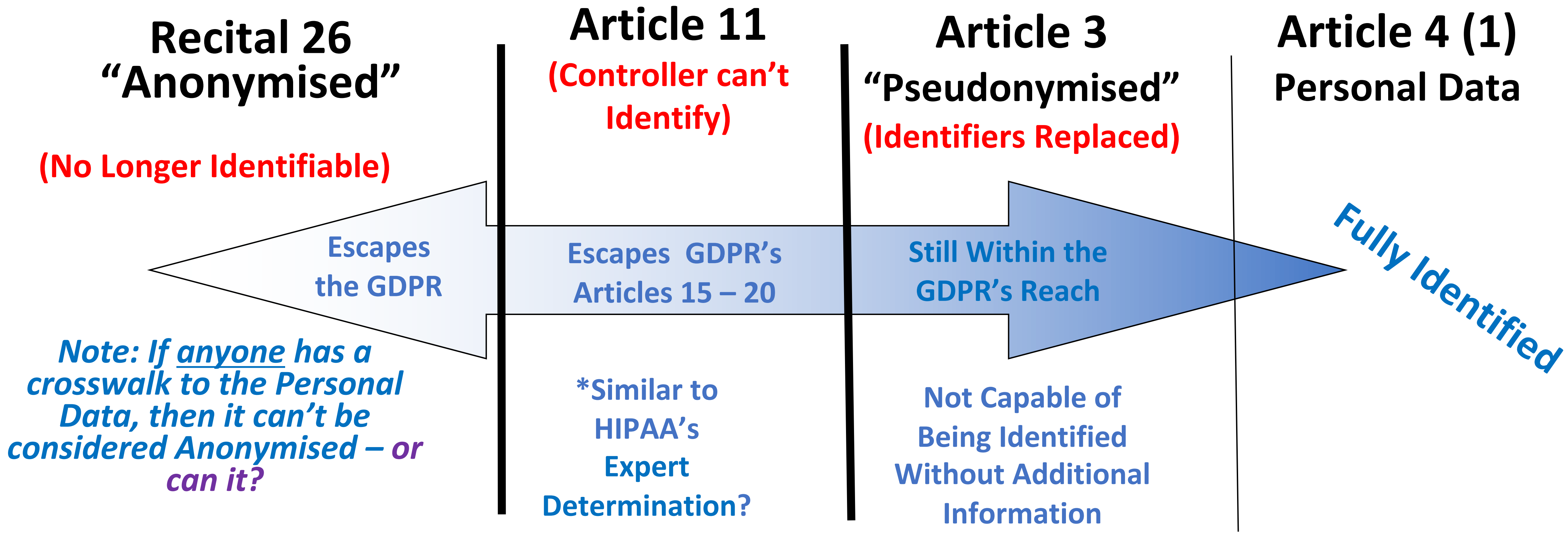
GDPR Background

- Definition of **personal data** ("any information relating to an identified or identifiable natural person");
- Definition of **pseudonymisation** ("the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject *without the use of additional information*, provided that such additional information is *kept separately* and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person");
- "Principles" for all processing of personal data (e.g., lawfulness, fairness, transparency; purpose limitation; data minimisation) (Art. 5);
- Legal bases for processing [purpose] (Art. 6 / Art. 9);
- No definition of anonymous data or anonymisation! Just **Recital 26**…

GDPR Background -- Recital 26 (*emphases added*)

The principles of data protection should apply to any information concerning an identified or identifiable natural person. [Personal data which have undergone **pseudonymisation**, which could be attributed to a natural person by the use of additional information should be considered to be information on an **identifiable** natural person.] [To determine **whether a natural person is identifiable**, account should be taken of **all the means reasonably likely to be used**, such as singling out, **either by the controller or by another person** to identify the natural person directly or indirectly.] [To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of **all objective factors**, such as **[1]** the costs of and the amount of time required for identification, **[2]** taking into consideration the available technology at the time of the processing and **[3]** technological developments.] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

GDPR’s Identification Risk/Legal Spectrum



Recital 26
“Anonymised”

(No Longer Identifiable)

Escapes the GDPR

Note: If <u>anyone</u> has a crosswalk to the Personal Data, then it can’t be considered Anonymised – or can it?

Article 11
(Controller can’t Identify)

Escapes GDPR’s Articles 15 – 20

***Similar to HIPAA’s Expert Determination?**

Article 3
“Pseudonymised”
(Identifiers Replaced)

Still Within the GDPR’s Reach

Not Capable of Being Identified Without Additional Information

Article 4 (1)
Personal Data

Fully Identified

*Hintze, Michael, *Viewing the GDPR through a De-Identification Lens: A Tool for Compliance, Clarification, and Consistency.* International Data Protection Law Vol 8, Iss 1, Feb 2018, Pgs 86–101, Available at https://ssrn.com/abstract=2909121

1st Threshold question: is there a requirement to have a legal basis to anonymise personal data?

- Put differently, does anonymisation constitute a purpose / processing of personal data that requires its own legal basis?
- If a separate legal basis is required, this might imply that only controller(s) can make that determination (controllers determine the *purpose* and means of processing of personal data).

[1st Threshold question] Support for “A separate legal basis is *not* required”

- “Anonymisation is not a purpose, it is a technique,” Irina Vasilou (DG Just/C3, EC; Workshop on GDPR Implementation and Health Data, Oct. 23, 2017) (in response to query as to whether consent forms should seek to cover anonymisation).
- Anonymisation is akin to erasure or destruction. The technical “how” of the processing of personal data is not itself a “purpose” (e.g., would not be set forth in a consent or notification). As a technique, is a “non-essential means”, which a controller can delegate to its processor. *See* Art. 29 WP Op. 01/2010 (“the determination of the “means” of processing can be delegated by the controller, as far as technical or organizational questions are concerned”).
- Data minimization principle.
- Requiring consent (or other, separate legal basis) for every ‘processing’ would seem to run counter to data protection principles. Recital 32, “Consent should cover all processing activities carried out for the same purpose or purposes”; *cf* .controller definition (determines the purpose and means).
- Plain text reading of the “compatibility test” does not equate any and all processing with a “purpose” for which a separate legal basis is required. Recital 50 (“The processing of personal data *for purposes* other than those for which the personal data were initially collected…”).

[1st Threshold question] Support for “A separate legal basis *is* required”

- WP Op. 05/2014 looked specifically at the “Lawfulness of the Anonymisation Process” and determined that a separate legal basis is required.
- May 2021 ICO draft guidance (Introduction to Anonymisation):

  “Techniques and approaches that are designed to turn personal data into anonymous information **constitute processing operations** performed on that data.”

  “This means that you need to comply with data protection requirements for this processing. This includes **ensuring you have a lawful basis** for it and you clearly define your purpose(s).”

  “In general it is likely that applying anonymisation techniques to the personal data you hold will be fair and lawful. However, it is still necessary for you to clearly **define your purpose** and detail the technical and organisational measures you intend to implement to achieve it.”
- German supervisory authorities, e.g., the Federal Commissioner for Data Protection and Freedom of Information (“BfDI”), June 2020 “Position Paper”.

→ ***It seems clear that a legal basis adheres to a purpose***, and not mere ‘processing’ or techniques. *See, e.g.*, EDPB Op. 03/2019 “concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR)”; Art. 6(4); Art. 9.

[1st Threshold question] What would that legal basis be?

- ***Compatibility***.
  - WP Op. 05/2014 concluded that “anonymisation as an instance of further processing of personal data can be considered to be compatible with the original purposes of the processing but only on condition the anonymisation process is such as to reliably produce anonymised information”.
  - GDPR prohibits secondary uses (“further processing”) without consent / a legal basis, unless (Art. 5):
    - that secondary use is **not “incompatible”** with the original “purpose” of the processing of that data
    - In addition, further processing for … “scientific [ ] research purposes” shall “**not** be considered incompatible with the initial purposes”, and so long as appropriate safeguards (TOMs) are in place (per Art. 89).
  - Art. 6(4) provides a (non-exhaustive) test for “ascertaining” whether that “another purpose” is compatible:
    - Any links between the purposes; the context of collection (data subjects and controller); the nature / sensitivity of the data; possible consequences for the data subjects; and the existence of safeguards (like pseudonymisation!)
- Compliance with a legal obligation (Art. 6(1)(c)).
- Right to Erasure (Art. 17 GDPR) (a specific legal obligation).

[2nd Threshold question] Does anonymisation need to be 'absolute'?

- Recital 26: "To determine whether a natural person is identifiable, account should be taken of ***all the means reasonably likely to be used***, such as singling out, either by the controller ***or*** by another person to identify the natural person directly or indirectly."
  - Note the "or"! In whose hands must the data be anonymised? Is this conjunctive or disjunctive?
  - Recital 26: "To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of **all objective factors**, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."
- Spain's DPA (Agencia Española de Protección de Datos) and the European Data Protection Supervisor releasing a joint document in April 2021 titled "10 misunderstandings related to anonymisation", which states:

  "Anonymisation is a process that **tries to find the right balance** between reducing the reidentification risk and keeping the utility of a dataset for the envisaged purpose"**.** "A robust anonymisation process **aims to reduce the re-identification risk below a certain threshold**." "Although a 100% anonymisation is the most desirable goal from a personal data protection perspective, in some cases it is not possible and a **residual risk of reidentification must be considered**."

[2nd Threshold question] Does anonymisation need to be ’absolute’ (cont’d)

- Despite apparent clarity based on GDPR and leading authorities:

> “[T]he concept of anonymization [has] remained clouded for decades, with EU data protection supervisory authorities and national courts holding anonymization is virtually impossible as long as someone, **even a third party**, can identify the respective person.”
>
> IAPP, “New Options for anonymization ahead?”, Ulrich Baumgartner (May 2023).

[3rd Threshold question] Does anonymisation depend on the ‘holder’?

- *Put differently, just how “absolute” does anonymisation have to be*? A shifting history:
- Article 29’s WP136 (2007) - a flexible, contextual standard
  - Whether data is personal data depends on “the specific scheme in which [a given] controller[] [is] operating, [e.g., if] reidentification is explicitly excluded and appropriate technical measures have been taken in this respect.”
- Article 29’s WP216 (2014) - a less flexible standard.
  - Pseudonymised data is necessarily personal data. “Pseudonymised data cannot be equated to anonymised information as they continue to allow an individual data subject to be singled out and linkable across different data sets. Pseudonymity is likely to allow for identifiability, and therefore stays inside the scope of the legal regime of data protection”
  - Anonymisation requires deletion or abstraction of the original data. “Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous”.

→ Not so helpful: anonymisation vs pseudonymisation is just a question of the likelihood of re-identifiability.
So why focus on categorizing pseudonymised data as necessarily personal data?

→ Not so odd: pseudonymisation in this context is really addressing the possibility of tracing / linking back.

[3rd Threshold question] Does anonymisation depend on the ‘holder’ (cont’d)

- Further guidance in WP216 (05/2014)
  - Protect against three specific re-identification risks:
    - Singling out an individual in a dataset;
    - Linking two records within a database; and
    - Inference (ability to deduce, with significant probability, information by using other information).
  - A de-identified (or otherwise anonymised) dataset remains personal data **where the controller retains the original** - from reading Recital 26’s “'the means likely reasonably to be used to determine whether a person is identifiable” as those used “by the controller [AND] by any other person” → conjunctive!
    - “An effective anonymisation solution **prevents all parties** from singling out an individual in a dataset, from linking two records within a dataset (or between two separate datasets) and from inferring any information in such dataset.”

Enter *SRB*: back to contextuality / relativity

- *SRB v. EDPS* (Case T-557/20, European General Court, April 26, 2023). Stands for the proposition, at least, that a dataset may be anonymous in the hands of its holder.
  - Not quite the same as 'pseudonymised Data is not personal data in the hands of a recipient that can't re-identify it'. At least one more step: the recipient must lack the "legal means" to re-identify the data.
  - *Background*: SRB pseudonymised survey responses by replacing participant names with randomly-generated alphanumeric codes, and shared this pseudonymised dataset (and not the decoding key) with a third party consultant. SRB did not inform the participants of this.
  - *Holding*: the EDPS failed to examine whether the authors of the comments were reidentifiable for Deloitte and whether such reidentification was reasonably possible.
  - *Rationale*: pseudonymised data in the hands of a third party is not personal data recipient where that recipient does not have the decoding information and no legal means of obtaining this information; the fact that the sender has the decoding key is irrelevant.
  - Tosses long-standing approach: "indirect" identifiability for the recipient would also be at hand if the identifying information (the alphanumeric code) is in the possession of another entity (the discloser).
  - EDPS is appealing this ruling to the Court of Justice of the European Union.

Enter *SRB*: What does “legal means” mean?

- Starting with Recital 26’s “all the means reasonably likely to be used”.
  - The holder does not have physical access to re-identifying data, and has no commercial, contractual, statutory or other legal right to obtain such access.
    - Contractual no re-identification provisions may help, but may be insufficient.
- *Breyer* test regarding whether a data holder has the “means likely reasonably to be used” to re-identify data in their possession:
  - Whether it is **lawful** for the holder to access the reidentifying data (i.e., “legal means”); and
  - Whether the holder is **reasonably likely** to gain physical access.
- Correcting the past: *Breyer* is not in fact ‘absolutist’; identifiability is entity-specific.
  - The ECJ in *Breyer* reviewed whether a dynamic IP address was personal data from the perspective of a website operator; it accepted without question that it was personal data for the internet service provider.
  - *SRB* may go further: the disclosed data is not personal data to the discloser (who did not therefore fail to provide information to data subjects).
  - *Breyer* is broadly over-read? Rather than an absolutist view, it clearly permits a residual risk of re-identification; each situation must be examined on its own merits.

Implications and choices; pseudonymisation.

- Implications: anonymised data in general; traceability.
  - Internal uses - for both controllers and processors (product improvement; product development; legitimate-interest-based marketing and non-marketing user segmentation; user experience improvements / driving adoptions; business insights).
  - Common business uses – vendors working with (anonymised) subsets of data (just as with *SRB*).
  - Life sciences - consented (prospective) research, let alone 'further research':
    - FDA draft guidance on external control arms ("Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products Guidance for Industry", Feb. 2023, here) would **require traceability** back to source data.
    - "Sponsors should also ensure that FDA has access to source documents and source data for the external control arm as part of an FDA inspection or upon request."
  - Effective route for anonymisation could help offset the "essential[] evisceration" of legal pathways to 'broad consent', e.g., consent for future, related research purposes that could not be identified at the time of consent. (WP259, April 2018, cf. Recital 33)
    - EDPB has since disparaged consent as legal (privacy) basis for processing personal data in clinical trials - finds a likely "power imbalance" between the investigator and data subject that prevents consent from being "freely given". (EDPB Op. 03/2019).

Implications and choices; pseudonymisation (cont'd)

- Attempt anonymisation; adopt a risk-based approach where not assured that the three risks are mitigated - this approach is possible per the (stricter) WP2014 guidance.
  - Challenges for intra-party work; maybe for internal use frameworks.
  - Would lean on attacker-centric approach; "assessment of the reidentification means reasonably likely to be used by the controller or another person, i.e, an attacker. In order to anticipate attackers' behavior, deidentification experts rely upon risk models to guide them in their selection of data and context controls."
- Maximize pseudonymisation / do not pursue anonymisation.
  - Pros: Recognized safeguard, including for transfers; most DSARs no longer apply; guidance on pseudonymisation exists, e.g., from "the European Union Agency for Cybersecurity"
  - Cons: Any party processing pseudonymised data for its own purpose would be a controller; reliance on 'compatibility' test?; treating information as personal data irrespective of its non-identifiability to its holder creates burden, delayed access and sharing across stakeholders.
- Trusted third-party approach
  - WP203 (April 2013) - achieve anonymisation by having a third party do the de-identification; first party retains the (identifiable) source data.
  - How does this (really) differ from *SRB* situation?

Implications and choices; pseudonymisation (cont'd)

- Technology-based solutions.
  - Synthetic (generative) datasets
    - Create anonymised data from source data by retaining and reflecting relationships / correlations in the source data, but none of that original data.
    - Medidata's generative modeling of synthetic clinical trial data (top award at the Int'l Conf on Machine Learning for Interpretable Machine Learning in Healthcare!) (U.S. Pat. No. 11,640,446)
      - Synthetic data preserves the structure and statistical correlations of the original dataset.
    - CNIL approval for "WeData". (The "National Commission for Information Technology and Liberties" certified that "avatar technology does not allow patients to be re-identified." "The data, once transformed, therefore no longer depends on the GDPR, since it is no longer personal data. There is no longer any link with the individual.")
  - Federated approaches, e.g., a distributed data network against which strictly anonymous summary information can be retrieved about member data sources; BYOM (ML models) can be applied to such sources; no transfer / source data remains local.

Implications and choices; pseudonymisation (cont'd)

- **Benefits of Pseudonymisation** (if anonymisation isn't possible, or even if it is!)
  - Synthetic (generative) datasets
    - Traceability / linkability - clear path, additional use cases (e.g., external control arm source data provenance guidance).
      - Though this would be against ENISA pseudonymisation recommendations! And would undercut planned / envisioned re-linking, e.g., key-coded clinical trial data 'emergency unblinding'.
      - And if *SRB* stands, perhaps this advantage goes away.
    - Potential for higher utility.
    - Enhanced security / compliance, with reduced DSAR obligations (Arts. 11(2), 12(2)).
    - Helps with EU to non-adequate country transfers: if recipient is not in a position to re-identify (lacks the additional information), is a "supplementary technical measure".
    - Is arguably very rigorous: **but for** the additional information that's held separately, no identifiability – no "reasonableness" standard baked in. The 'content' itself has to be non-identifiable, i.e. anonymous!

Implications and choices; pseudonymisation (cont'd)

- *How much re-identification risk is acceptable*?
  - One domain with a standard: clinical trial research.
  - EMA's Policy 0070 seeks to publish information derived from clinical trials data.
  - As the EMA is subject to the GDPR's requirements, its anonymisation standard a proxy for a standard under GDPR. Per its "External Guidance on the Implementation of Policy 0070" (Sept. 20, 2017), "EMA believes that it is advisable to set the threshold to a conservative level of **0.09**."
- Short of a standard set by an authority, may controllers set their own standards?
  - May the European Commission? Consider the draft "European Health Data Space".

Questions + Contact

Daniel Barth-Jones, PhD

Privacy Expert in Residence
Privacy Hub by Datavant

danielbarth-jones.privacyhub
@datavant.com

Ann Waldo, JD

Waldo Law Offices

awaldo@waldolawoffices.
com

Andrew Kopelman

Chief Privacy Counsel
Medidata Solutions
dataprivacy@mdsol.com;
Andrew.Kopelman@3ds.com

Claire Manneh, MPH

Head of Provider Partnerships
Datavant

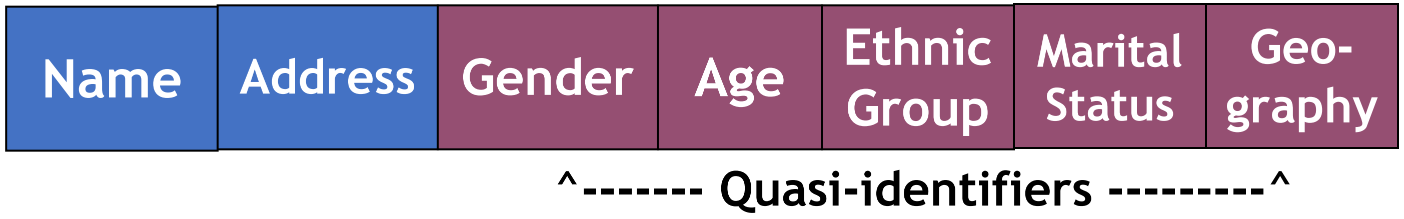
claire@datavant.com

Reference Slides

More nuances re: de-identification

Quasi-identifiers

While individual fields may not be identifying by themselves, the contents of several fields in combination may be sufficient to result in identification, the set of fields in the Key is called the set of ***Quasi-identifiers***.



Name	Address	Gender	Age	Ethnic Group	Marital Status	Geo-graphy

^------- Quasi-identifiers ---------^

Fields that should be considered part of the Quasi-identifiers are those variables which would be likely to exist in “reasonably available” data sets along with actual identifiers (names, etc.).

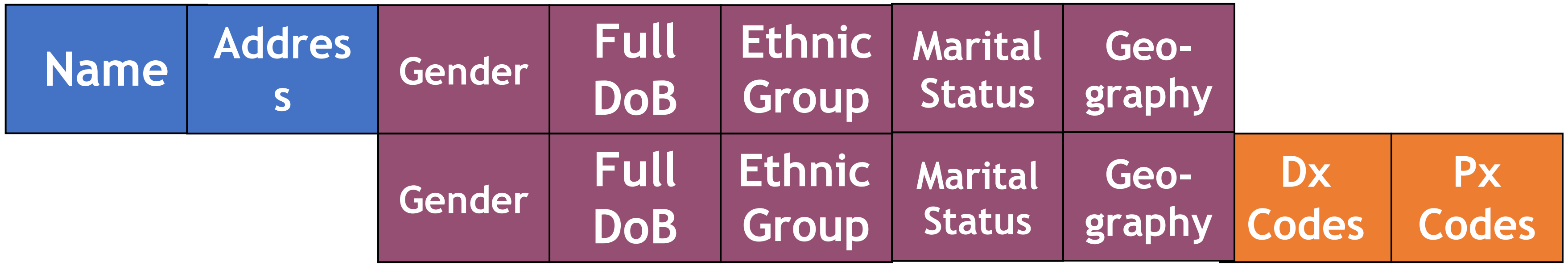
Note that this includes even fields that are not “PHI”.

Key Resolution

Key "*resolution*" exponentially increases with:

1) the number of matching fields available

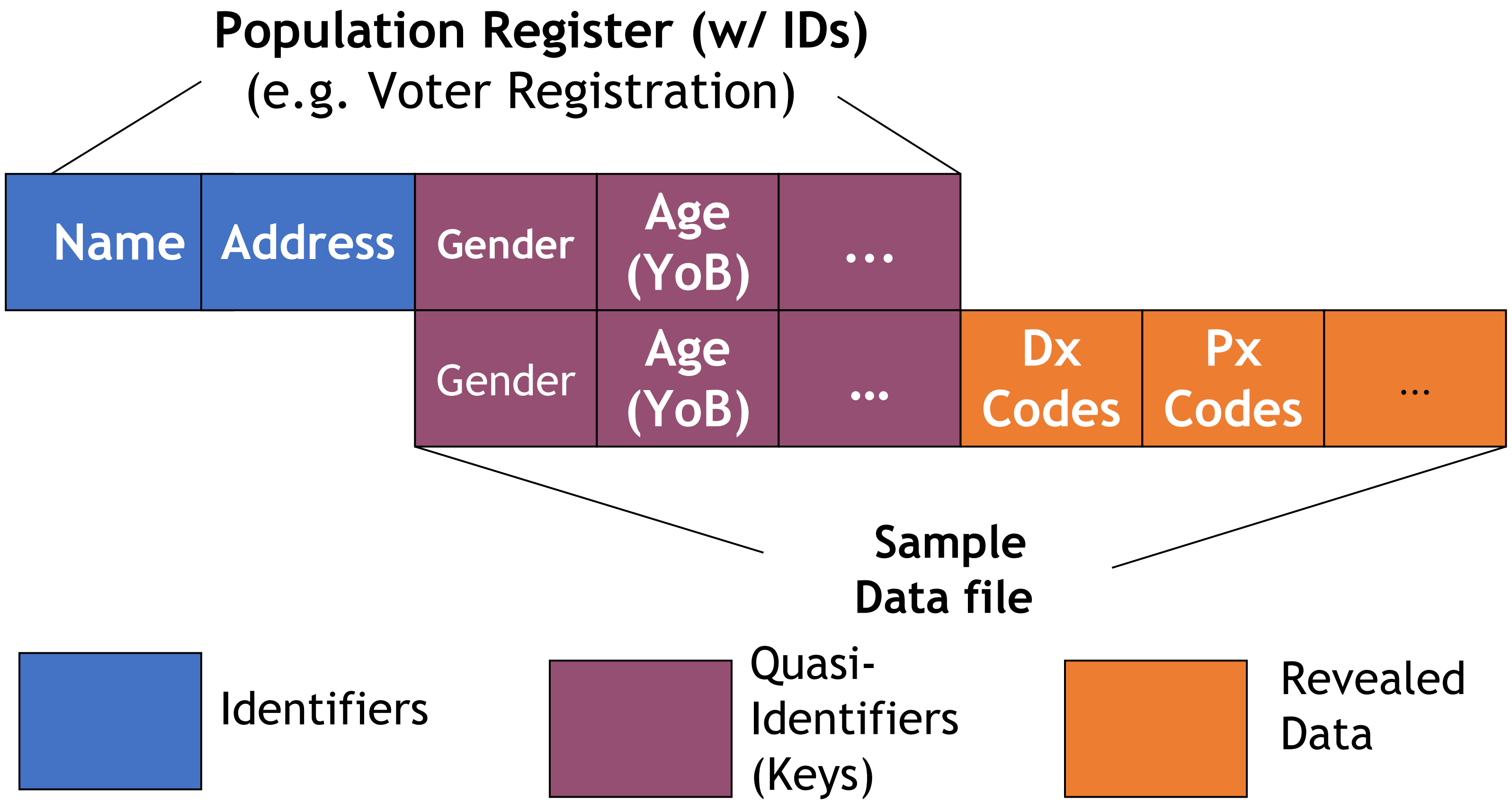
1) the level of detail within these fields. (e.g. Age in Years versus complete Birth Date: Month, Day, Year)



Name	Address	Gender	Full DoB	Ethnic Group	Marital Status	Geography		
		Gender	Full DoB	Ethnic Group	Marital Status	Geography	Dx Codes	Px Codes

Record Linkage

Record Linkage is achieved by matching records in separate data sets that have a common “Key” or set of data fields.



Population Register (w/ IDs)
(e.g. Voter Registration)

Name | Address | Gender | Age (YoB) | …

Gender | Age (YoB) | … | Dx Codes | Px Codes | …

Sample Data file

Identifiers

Quasi-Identifiers (Keys)

Revealed Data

Balancing Disclosure Risk/Statistical Accuracy

- Balancing disclosure risks and statistical accuracy is essential because some popular de-identification methods (e.g. k-anonymity, noise injection) can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses or data mining (distorting variance-covariance matrices, masking heterogeneous sub-groups which have been collapsed in generalization protections)
- This problem is well-understood by statisticians, but not as well recognized and integrated within public policy.
- Poorly conducted de-identification can lead to "bad science" and "bad decisions".

Reference: C. Aggarwal http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf

HIPAA §164.514(b)(1)(i) and *Anticipated Recipients*

(i) Applying such principles and methods, determines that the *risk is very small* that *the information could be used*, alone or *in combination with other reasonably available information, by an anticipated recipient to identify an individual* who is a subject of the information;

It is important to note that §164.514(b)(1)(i) is written with respect to "Anticipated Recipients". This introduces the concept of using policy, procedural and contract controls for limiting the Anticipated Recipients and the time periods and projects for which data is made available.

(See Q2.8., 2012 HHS De-identification Guidance pg. 18)

Ethical Equipoise?

Is it an ethically compromised position, in the coming age of personalized medicine, if we end up purposefully masking the racial, ethnic or other groups (e.g. American Indians or LDS Church members, etc.), or for those with certain rare genetic diseases/disorders, in order to protect them against supposed re-identification, and thus also deny them the benefits of research conducted with de-identified data that may help address their health disparities, find cures for their rare diseases, or facilitate "orphan drug" research that would otherwise not be economically viable, especially if those re-identification attempts may not be forthcoming in the real-world?

HHS Guidance (Nov 26, 2012)
Q2.2 ”Who is an “expert?” *(p. 10)*

- No specific professional degree or certification for de-identification experts.
- Relevant expertise may be gained through various routes of education and experience.
- Experts may be found in the statistical, mathematical, or other scientific domains.
- From an enforcement perspective, OCR would review the relevant professional experience and academic or other training of the expert, as well as their actual experience using health information de-identification methodologies.

HHS Guidance

Q2.3 *Acceptable level of identification risk?* *(p.11)*

- There is no explicit numerical level of identification risk that is deemed to universally meet the “very small” level.
- The ability of a recipient of information to identify an individual is dependent on many factors, which an expert will need to take into account while assessing the risk.

HHS Guidance
Q2.4 How long is an expert determination valid? *(p.11)*

- The Privacy Rule does not explicitly require an expiration date for de-identification determinations.
- However, experts have recognized that technology, social conditions, and the availability of information change over time. Consequently, certain de-identification practitioners use the approach of time-limited certifications.
- The expert will assess the expected change of computational capability and access to various data sources, and determine an appropriate time frame.

Q2.5 ***Can an expert derive multiple solutions from the same data set for a recipient?*** ***(p.11)***

- Yes. Experts may design multiple solutions, each of which is tailored to the information reasonably available to the anticipated recipient of the data set.
- The expert must take care to ensure that the data sets cannot be combined to compromise the protections.
  - Example: An expert may derive one data set with detailed geocodes and generalized age (e.g., 5-year age ranges) and another data set that contains generalized geocodes (e.g., only the first two digits) and fine-grained age (e.g., days from birth).

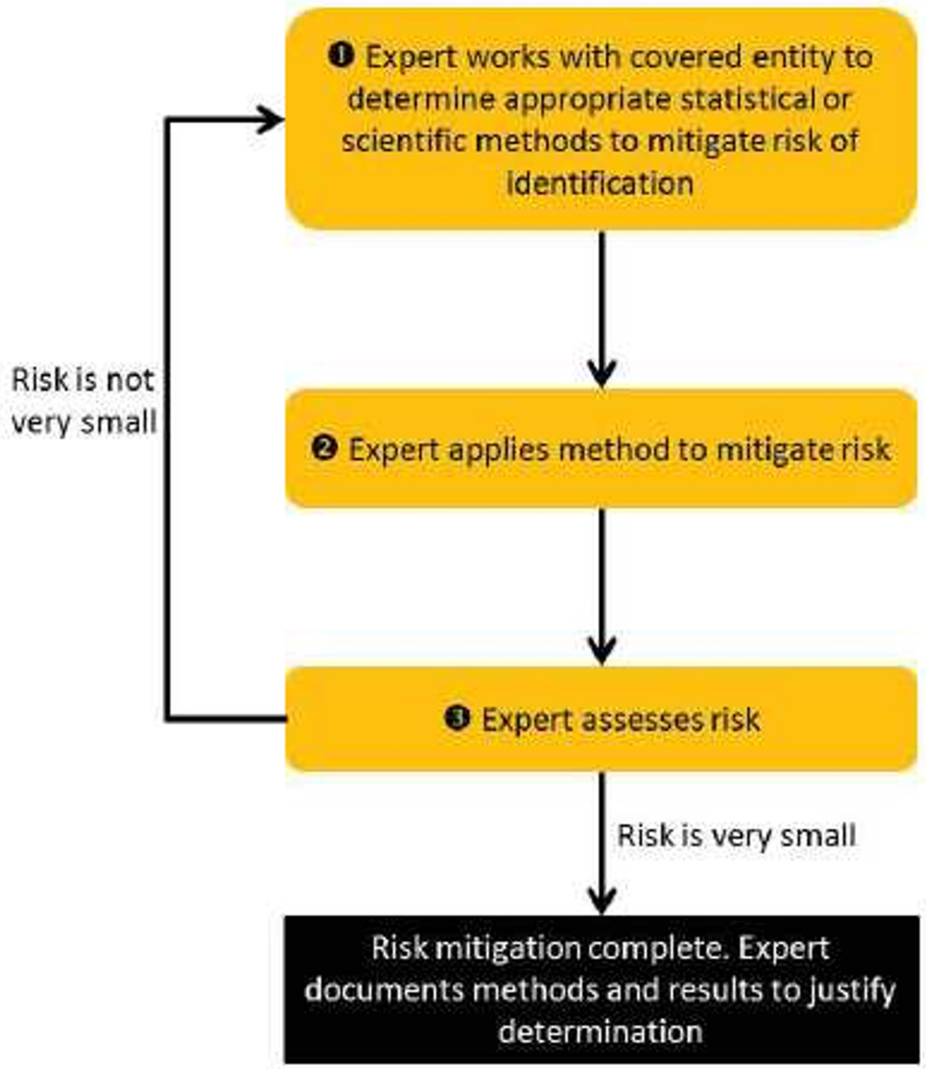
Q2.5 ***Can an expert derive multiple solutions from the same data set for a recipient?*** ***(Cont'd)***

- The expert may certify both data sets after determining that the two data sets could not be merged to individually identify a patient.
- This determination may be based on a technical proof regarding the inability to merge such data sets.
- Alternatively, the expert also could require additional safeguards through a data use agreement.

Q2.6. *How do experts assess the risk of identification of information?* *(p.12-16)*

- No single universal solution
- A combination of technical and policy procedures are often applied.
- OCR does not require a particular process for an expert to use to reach a determination that the risk of identification is very small.
- The Rule does require that the methods and results of the analysis that justify the determination be documented and made available to OCR upon request.

General Workflow for Expert Determination



❶ Expert works with covered entity to determine appropriate statistical or scientific methods to mitigate risk of identification

❷ Expert applies method to mitigate risk

❸ Expert assesses risk

Risk is not very small

Risk is very small

Risk mitigation complete. Expert documents methods and results to justify determination

The De-identification process may require several iterations until the expert and data managers agree upon an acceptable solution.

Figure 2. Process for expert determination of de-Identification.

Q2.8. *What are the approaches by which an expert mitigates the risk of identification?* *(p.18)*

- The Privacy Rule does not require a particular approach to reduce the re-identification risk to very small.
- In general, the expert will adjust certain features or values in the data to ensure that unique, identifiable elements are not expected to exist.
- An overarching common goal of such approaches is to balance disclosure risk against data utility.

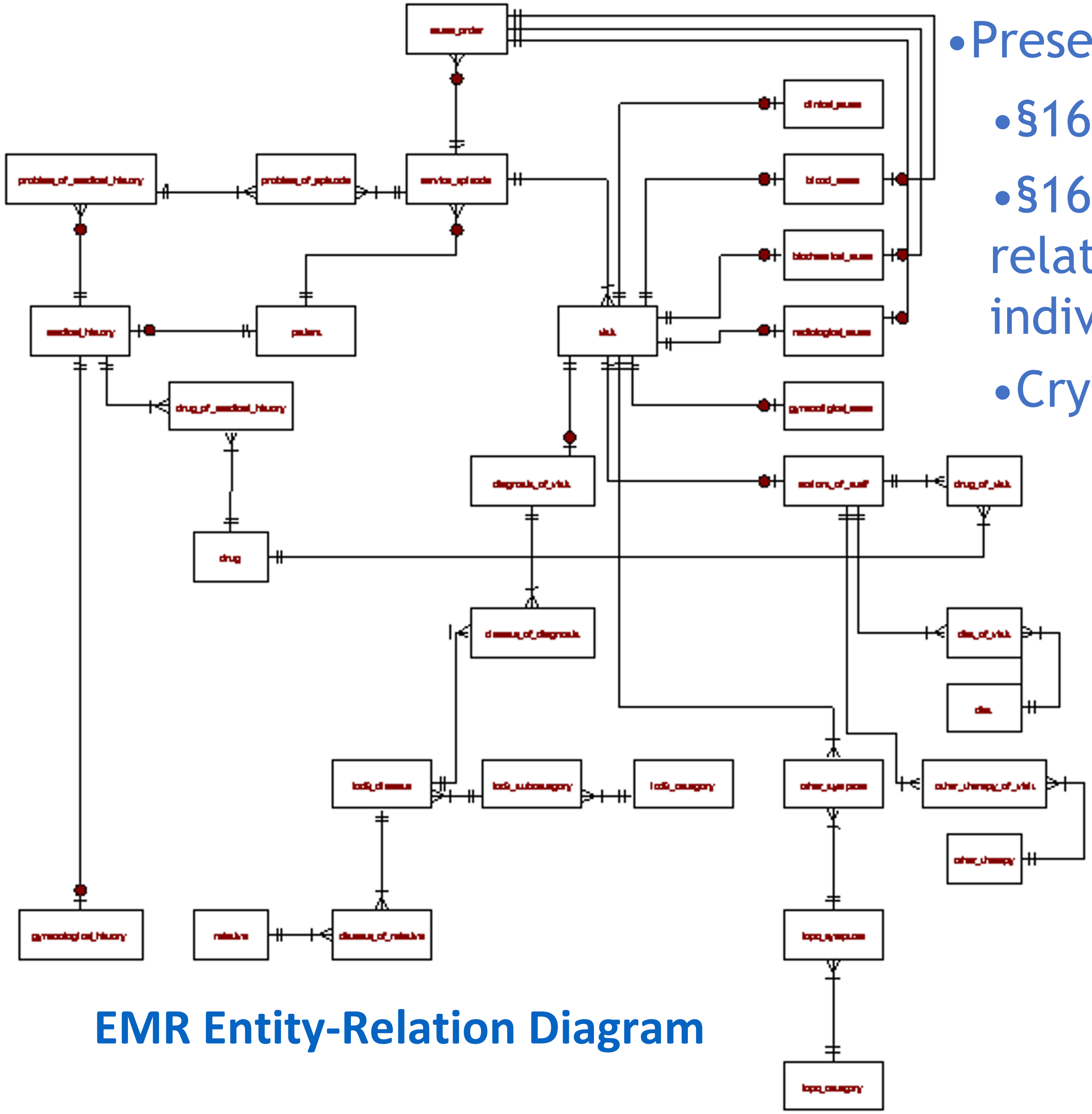
Q2.8. ***What are the approaches by which an expert mitigates the risk of identification?*** ***(Cont'd)***

- Determination of which method is most appropriate will be assessed by the expert on a case-by-case basis.
- The expert may also consider limiting distribution of records through a data use agreement or restricted access agreement in which the recipient agrees to limits on who can use or receive the data, or agrees not to attempt identification of the subjects. Specific details of such an agreement are left to the discretion of the expert and covered entity.

Q2.9 ***Can an Expert determine a code derived from PHI is de-identified?*** *(p.21-22)*

- A common de-identification technique for obscuring information is to use a one-way cryptographic function (known as a hash function)
- Disclosure of codes derived from PHI in a de-identified data set is allowed if an expert determines that the data meets the requirements at §164.514(b)(1). The re-identification provision in §164.514(c) does not preclude the transformation of PHI into values derived by cryptographic hash functions using the expert determination method, provided the keys associated with such functions are not disclosed.

Complexities for Longitudinal De-identification



EMR Entity-Relation Diagram

- Preserving Referential Integrity
  - §164.514(b)(2)(i)(R): Unique code exclusion
  - §164.514(c)(1): Not “derived from or related to information about the individual”
  - Cryptographic Hashing Solutions
    - Correctly identifying and de-identifying patients across repeated encounters

Audience Question

If you have a national dataset, which state laws apply?

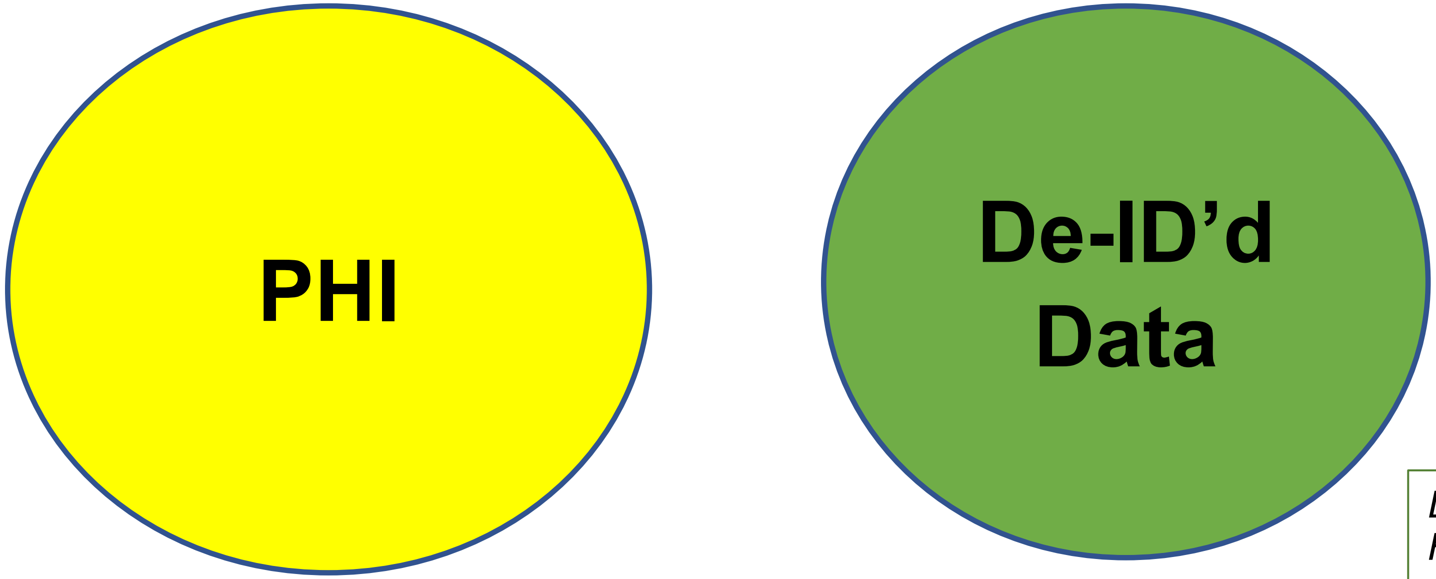
Put differently, what is the jurisdictional hook for each state law?

Audience Question

Which de-ID'n standard do you think applies if PHI is combined with consumer data prior to de-ID'n?

De-identification under HIPAA - Basics

Sharp legal divide in HIPAA between de-identified data and PHI



PHI

De-ID’d Data

De-ID’d data is outside HIPAA
HHS has no jurisdiction
Contract restrictions may apply

Federal ADPPA - Another De-ID’n definition AND No HIPAA Harmonization

DE-IDENTIFIED DATA.— The term “de-identified data” means information that does not identify and is not linked or reasonably linkable to a distinct individual or a device, regardless of whether the information is aggregated, and if the covered entity or service provider—

(A) takes reasonable technical measures to ensure that the information cannot, at any point, be used to re-identify any individual or device that identifies or is linked or reasonably linkable to an individual;

(B) publicly commits in a clear and conspicuous manner—

(i) to process and transfer the information solely in a de-identified form without any reasonable means for re-identification; and

(ii) to not attempt to re-identify the information with any individual or device that identifies or is linked or reasonably linkable to an individual; and

(C) contractually obligates any person or entity that receives the information from the covered entity or service provider—

(i) to comply with all of the provisions of this paragraph with respect to the information; and

(ii) to require that such contractual obligations be included contractually in all subsequent instances for which the data may be received.