May 8, 2024

De-identification Workshop

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

Ann Waldo, JD
Principal
Waldo Law Offices

David Copeland, PhD
Senior Data Scientist and Privacy Expert
Privacy Hub by Datavant

Chris Diaz, JD
Senior Dir., Associate General Counsel – Privacy & AI (DPO)
Medidata Solutions, Inc.

Privacy+ Security Forum

Speaker

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

David Copeland, PhD
Senior Data Scientist and Privacy Expert
Privacy Hub by Datavant

David has been a data scientist at Privacy Hub and its legacy company Mirador Analytics since June 2021. He has performed statistical disclosure risk analyses on a wide range of real-world datasets, specializing in Unstructured Data such as free-text, images and genomics. He has developed several products and initiatives aimed at achieving the optimal balance of privacy and utility across these data types. David has co-authored government RFI's, provided commercial privacy training, and attended multiple privacy conferences. Prior to joining Privacy Hub, David was a postdoctoral research scientist at the Institute of Astronomy at the University of Edinburgh where he had previously received his Ph.D. in Astronomy.

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, de-identification, and innovation.

Speaker

Chris Diaz, CIPM, CIPP/US
Senior Director, Associate General Counsel – Privacy & AI (DPO)
Medidata Solutions, Inc. – a Dassault Systemes Company

Chris currently serves as the DPO for Medidata Solutions, Inc., a leading mission-driven life sciences SaaS platform that continues to transform and improve the conduct of global clinical trials. He manages the 'Privacy & AI' arm of the Medidata legal team, which advises both internal and external stakeholders on the wide range of data privacy and AI issues arising during the life cycle of a clinical trial. Immediately prior to Medidata Solutions, Inc., Chris was privacy counsel at Tapestry, Inc., moving in-house after serving as principal of his own practice. He received his J.D. from the University of California, Berkeley and his B.A. from the University of California, Los Angeles. Chris started his legal career at Ropes & Gray LLP.

Chris is currently a co-chair of the IAPP KnowledgeNet Chapter for New York and is a member of the Global Data Protection and Privacy Committee for the Association of Clinical Research Organizations (ACRO). On his spare time, he enjoys volunteering with his local animal rescue and appreciating the great outdoors.

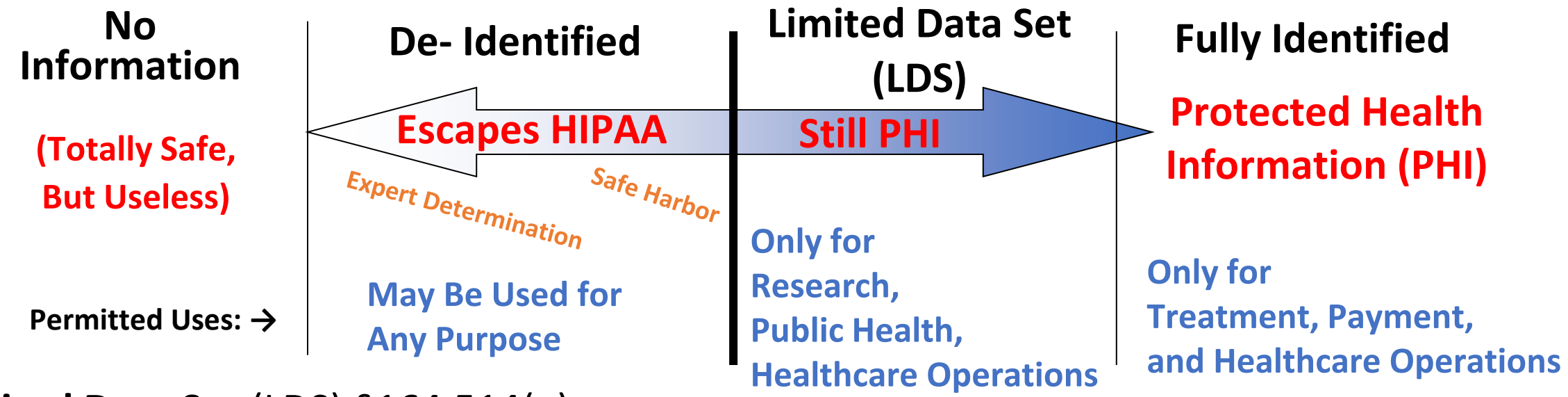
Overall Workshop Questions

- **What is de-identification – under HIPAA, EU law, and evolving state laws?**
- **What are the statistical, technical, and privacy-preserving challenges? What about emerging areas like images, unstructured data, synthetic data, and genomic data?**
- **Why does de-identification matter in the real world? What can de-identified data accomplish?**
- **What new technologies can make it more viable to extract scientific insights from linked de-identified data ?**
- **How might AI affect de-identification (for good or ill)?**
- **How have the new de-ID'n definitions in the new state laws changed things?**
- **What can organizations do to manage divergent de-ID definitions?**
- **What new state law obligations attach to de-ID'd data?**
- **What would the new federal bill say about de-ID?**
- **What's the latest on pseudonymization/anonymization in EU?**
- **Any reasons to hope for clarity around anonymisation under GDPR?**

Framing De-Identification

Daniel Barth-Jones

HIPAA's Identification Risk/Legal Spectrum



Limited Data Set (LDS) §164.514(e)

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

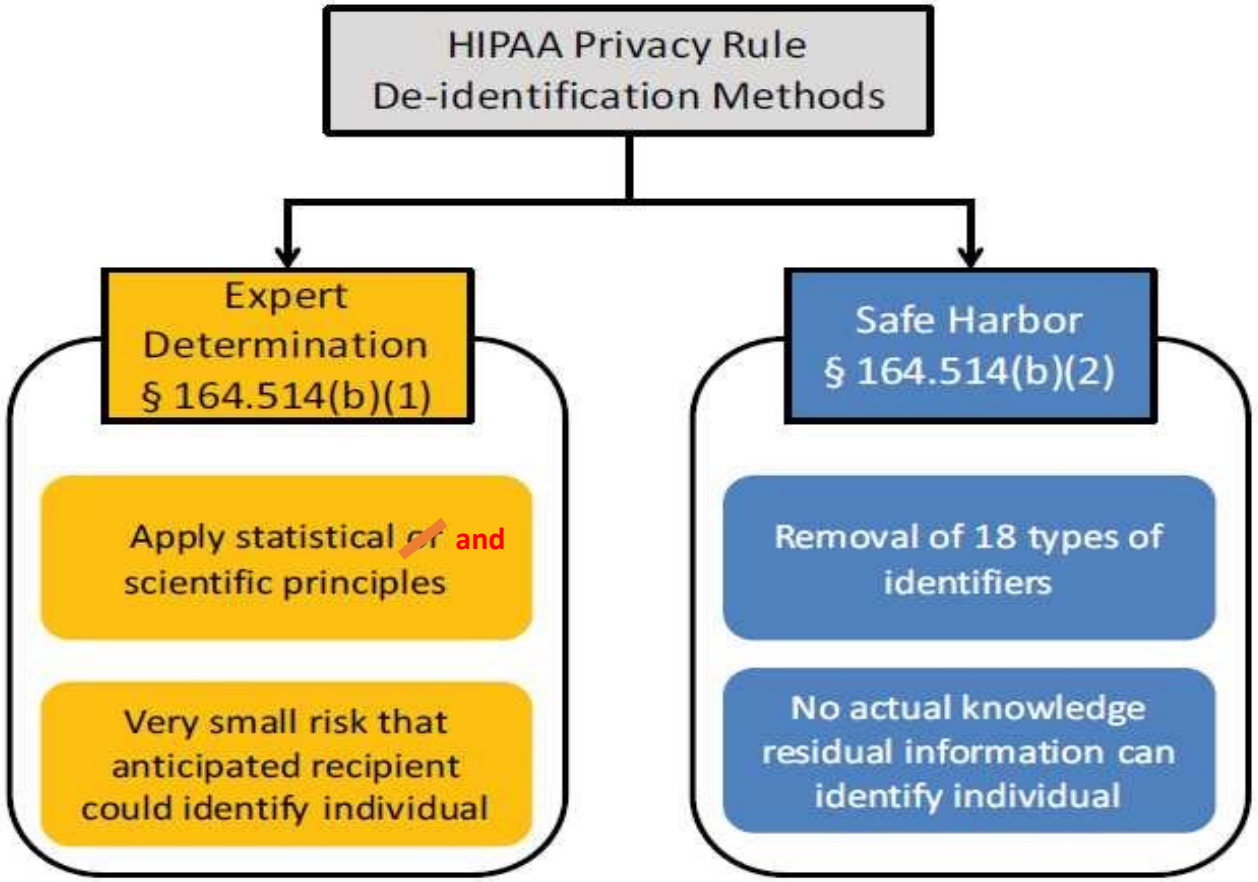
Safe Harbor De-identified §164.514(b)(2)

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



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"

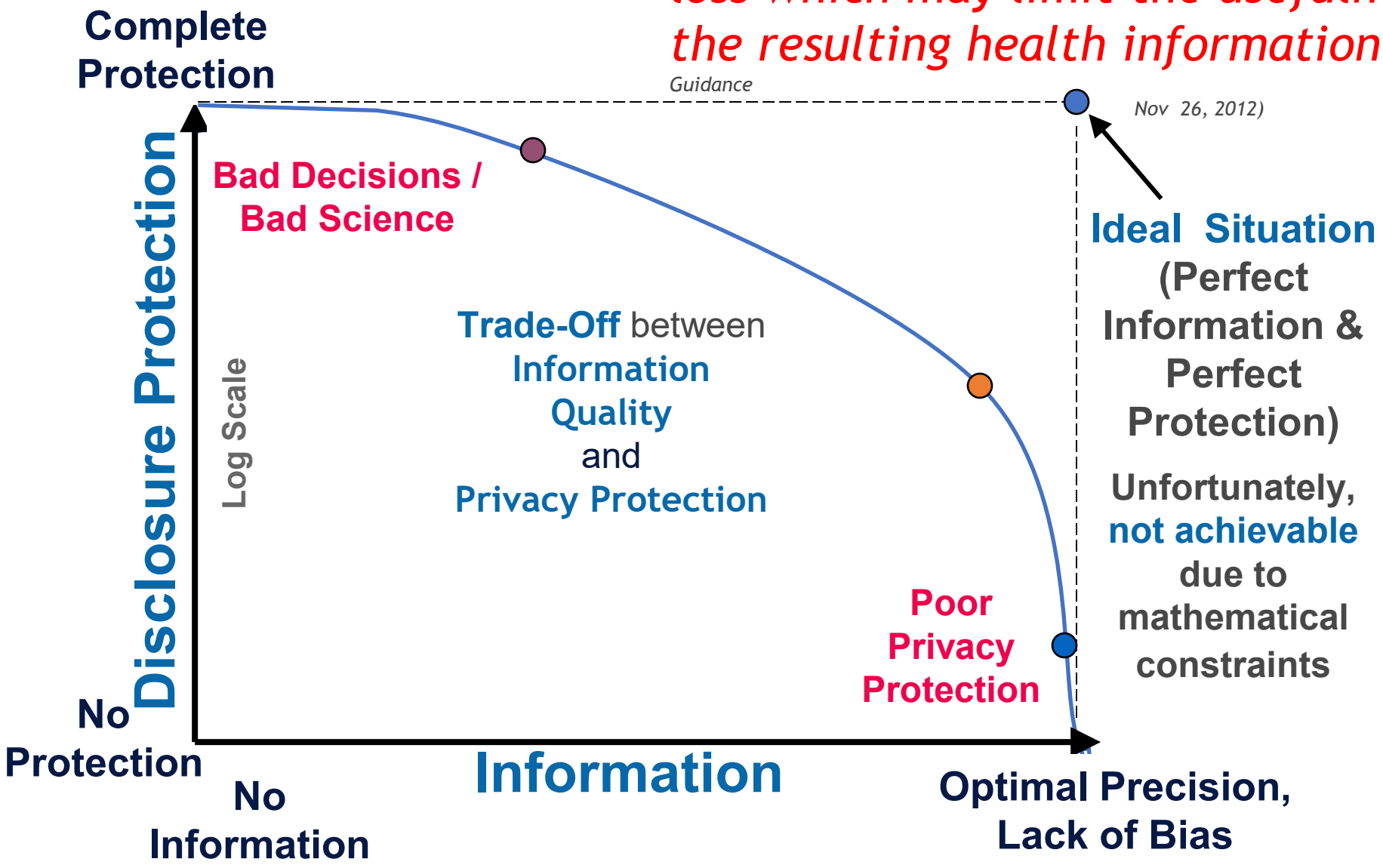
Health Information is not individually identifiable if:

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

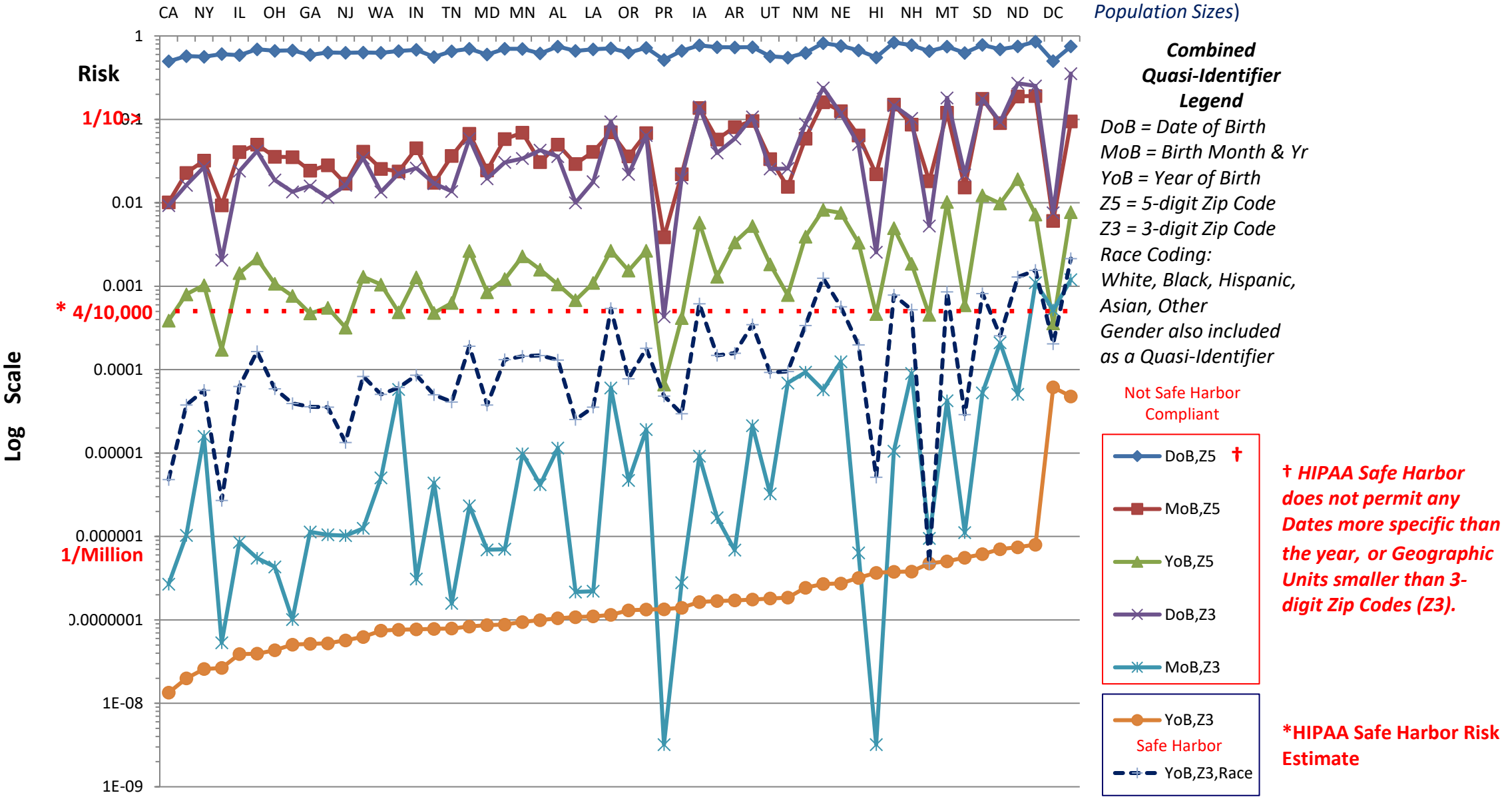
(i) Applying such principles and methods, determines that the *risk is very small* that *the information could be used*, alone or *in combination with other reasonably available information, by an anticipated recipient to identify an individual* who is a subject of the information; and (ii) Documents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



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



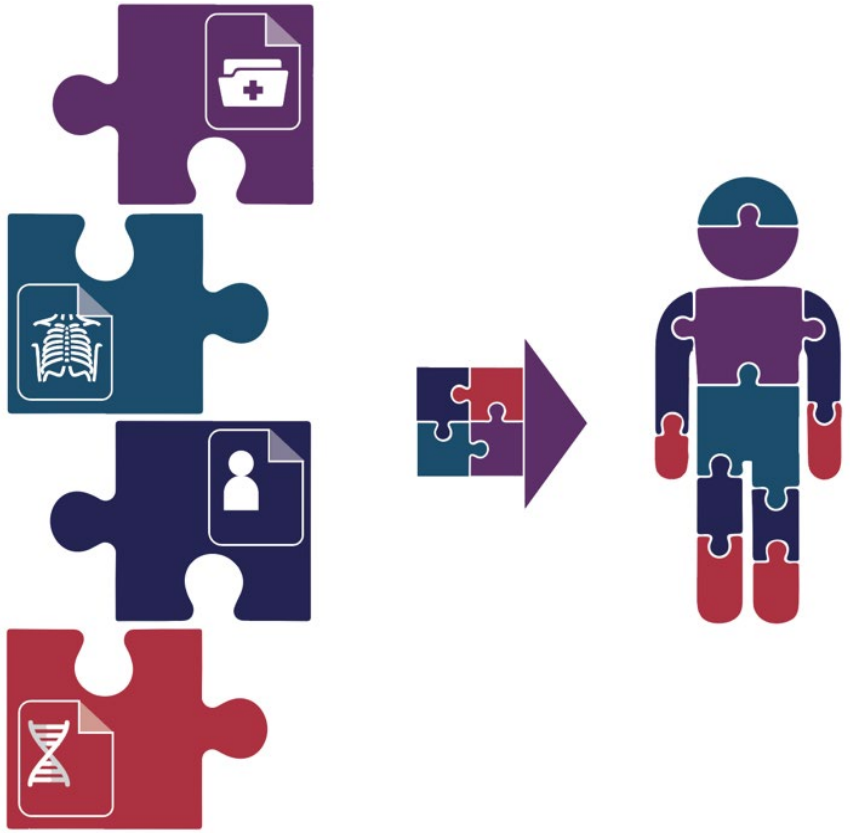
Data Source: 2010 U.S. Decennial Census

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

The Privacy Challenge of "*Putting the Patient Back Together*"



HIPAA Records Linkage Challenges/Solutions

- HIPAA prohibits the sharing of Protected Health Information (PHI) outside of established legal pathways (TPO, public health, etc.)
- 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 sources
- But there is a crucial need in nearly all 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's identity

Connecting health data manually is time and effort intensive and subject to multiple friction points

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

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 123

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

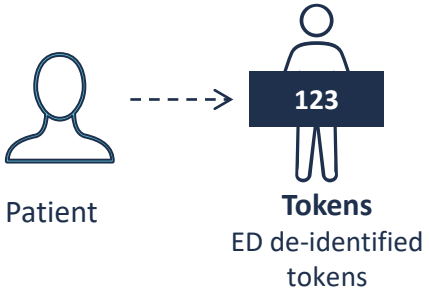
Patient 456

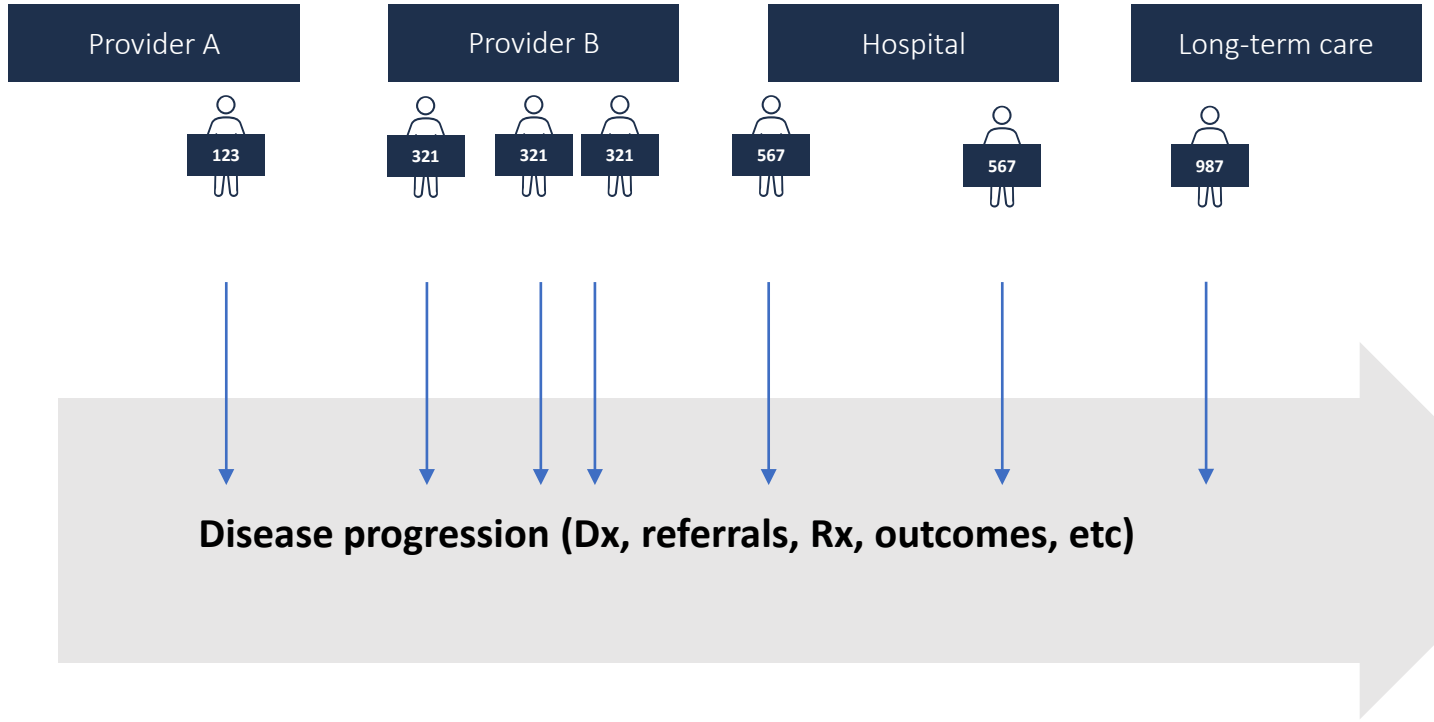
Patient Key:
Patient 123

Patient Key:
Patient 456

PPRL matches Patient 123 and Patient 346
as the same patient to get the complete picture

Tokenized IDs can allow linking of a patient's records across multiple sources to build a longitudinal view of their journey and aspects of their care





De-identified view of the patient's journey derived from multiple EHRs

Tokenization Process Example

	Input data		Data normalization*	Remediation	Output data	
	Dataset A	Dataset B			Dataset A	Dataset B
First Name	Abby	Abigail	Nickname resolution	Remove	Null	Null
Last Name	O'Grady, MD.	OGrady	Suffix stripped Special characters removed	Remove	Null	Null
Gender	Female	Female	No action	Pass through	Female	Female
Date of Birth	June 5, 1987	June 5, 1987	No action	Convert to birth year	1/1/1987	1/1/1987
Address	5 Pine Street	Five Pine Street	Spelling changed to number	Remove	Null	Null
Zip Code	95432	95432	No action	Convert to 3-digit zip	954	954

Encrypted keys and unique IDs created (from name, DOB and gender)

Patient Tokens

AA001 AA001

Tokens Created with Keyed ("Salted") One-Way Cryptographic Hash of Concatenated, Normalized Inputs

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

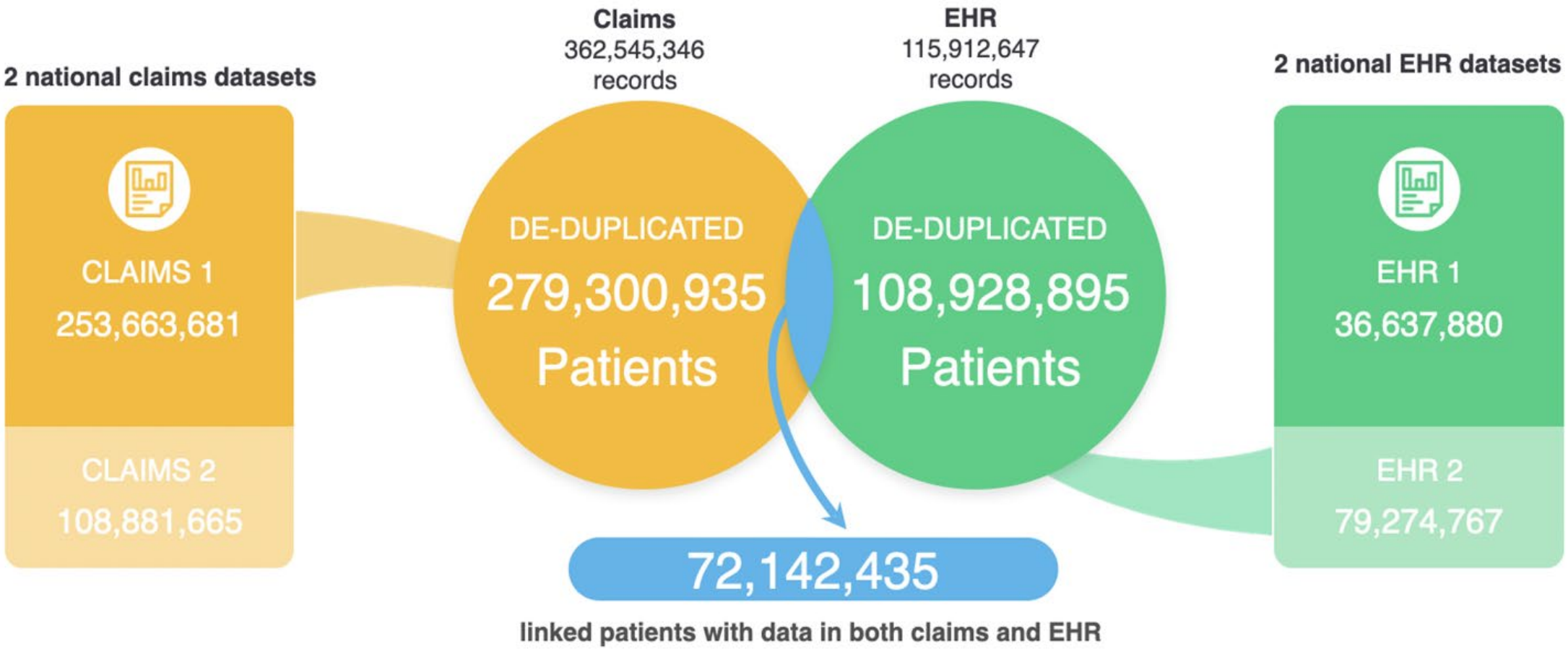
Collection of study data from 1m+ people in the US

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



Journal of the American Medical Informatics Association, 28(3), 2021, 427–443
doi: 10.1093/jamia/ocaa196
Advance Access Publication Date: 17 August 2020
Research and Applications

Research and Applications

The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment

Melissa A. Haendel [1,2], Christopher G. Chute [3], Tellen D. Bennett [4], Eichmann [5], Justin Guinney [6], Warren A. Kibbe [7], Philip R.O. Payne

Hill et al. BMC Public Health (2023) 23:2103
https://doi.org/10.1186/s12889-023-16916-w

BMC Public Health

RESEARCH ARTICLE Open Access

Risk factors associated with post-acute sequelae of SARS-CoV-2: an N3C and NIH RECOVER study

...ngh[5], Catherine Xie[6],

Journal of Clinical and Translational Science

www.cambridge.org/cts

Translational Research, Design and Analysis Special Communication

Cite this article: Suver C, Harper J, Loomba J,

The N3C governance ecosystem: A model socio-technical partnership for the future of collaborative analytics at scale

Christine Suver[1], Jeremy Harper[2], Johanna Loomba[3], Mary Saltz[4], Julian Solway[5], Alfred Jerrod Anzalone[6], Kellie Walters[7], Emily Pfaff[7], Anita Walden[8], Julie McMurry[8], Christopher G. Chute[9] and Melissa Haendel[8], on behalf of the N3C Consortium

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

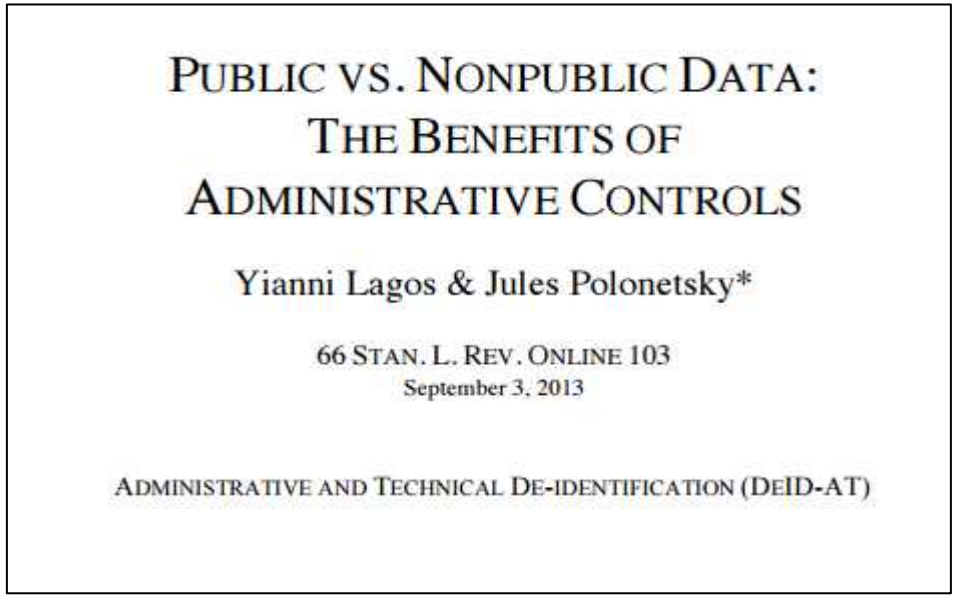
1,024 community clinics

3,564 primary care practices

338 emergency departments

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) –

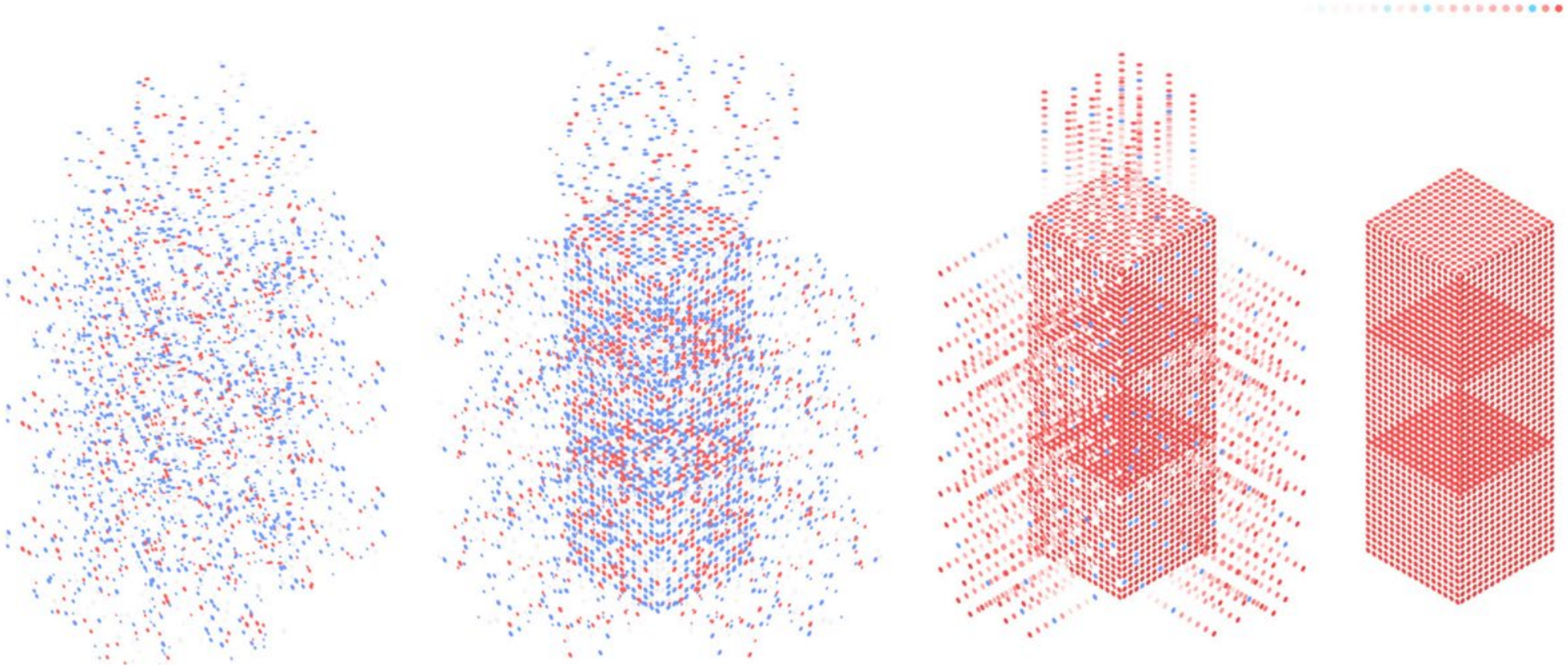
Which bans re-identification of previously de-identified health data, ***except*** *where such re-identification is* ***needed for HIPAA-governed activities****, is required by law, or* ***where necessary for testing, analysis, or validation of de-identification techniques.***

Should it be applied nationally?

Emerging Privacy – Unstructured Data and AI

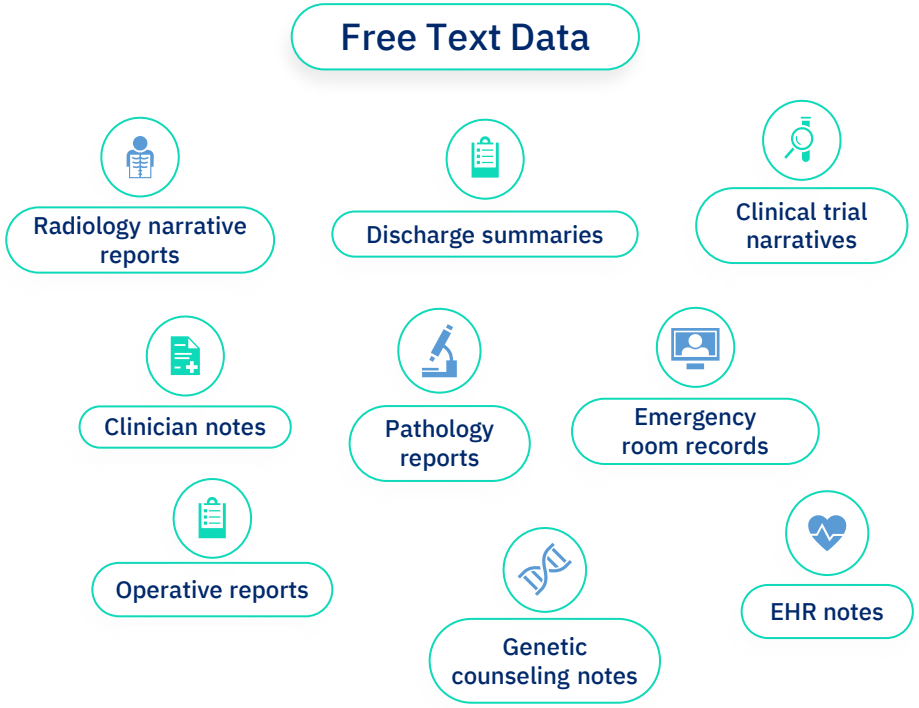
David Copeland

Unstructured Data



Free-text is a significant portion of the world's health data

Unstructured text is critical to understand the complete patient experience...



...but it is challenging to access and analyze while preserving patient privacy.

To derive value from free text datasets,
achieving legal de-identification is essential.

Free-text: where is the risk?

Lab notes e.g.“18mg on **12/19**”

Physician comments e.g. “**Mr Brown** had a heart attack and **passed away**.”

Multi-page documents e.g. a radiology report: **potentially any identifying information requiring manual review to verify absence/presence**

PatientID, Notes, Notes
Surgery_a4, “
name: Ben Stinson
PT ID: 6843115 65
DOB: 10/05/1940

Home Phone:(385) 3528050
Address: 979 Atrium street, Greendale, Bargersville, NE 40065
Insurance No: 85 0348

Date of visit: 25-Aug-63

Huntington Hospital
Address: 100 Ws California Blvd, Pasadena, CA 91105 Drs D Pateniak and vs Naipaul

PATIENT SURGICAL HISTORY

assessment requested by ins company (JAYA INT Ltd).
On 11-Aug 7:00 a.m. Ms. Stinson. was admitted to Huntington Hospital and under the care and treatment of Surgeon JJ Waters (General Surgeon). At 10:22 a.m. she was taken to the Operating Room holding area and beginning at 12:22 p.m. she underwent a Total Thyroidectomy for a malignant tumor. General Anesthesia was administered and monitored by Dr. Andrews, Anesthesiologist. Fentanyl, Atropine and Droperidol was given.
During the procedure Dr. Waters noted the mass to be "large" and "infiltrating". Frozen Section biopsy revealed Papillary Adenocarcinoma and a 3 ½ hour, Thyroidectomy was completed. Ms. Stinson was taken to the PACU for recovery at 2:45 p.m. and noted in stable condition however, her blood pressure was elevated reading 220/92. (Nurse George Fredrick, RN:4985)

signed: ETK Delafield MD
22-Dec-87”,

PatientID, Notes, Notes
Surgery_a4, “
name: **NAME**
PT ID: **PT ID**
DOB: **DOB**

Home Phone: **PHONE**
Address: **PT ADDRESS**
Insurance No: **INSURANCE ID**

Date of visit: **DATE**

Huntington Hospital
Address: **PROVIDER ADDRESS**

PATIENT SURGICAL HISTORY

assessment requested by ins company (JAYA INT Ltd).
On **DATE NAME** was admitted to Huntington Hospital and under the care and treatment of Surgeon **NAME** (General Surgeon). At **TIME** she was taken to the Operating Room holding area and beginning at **TIME** she underwent a Total Thyroidectomy for a malignant tumor. General Anesthesia was administered and monitored by **NAME**, Anesthesiologist. Fentanyl, Atropine and Droperidol was given.
During the procedure **NAME** noted the mass to be "large" and "infiltrating". Frozen Section biopsy revealed Papillary Adenocarcinoma and a **TIME**, Thyroidectomy was completed. **NAME** was taken to the PACU for recovery at **TIME** and noted in stable condition however, her blood pressure was elevated reading 220/92. **NAME, PROVIDER ID**)

signed: **NAME**
DATE,

Free-text: Paths to de-identification

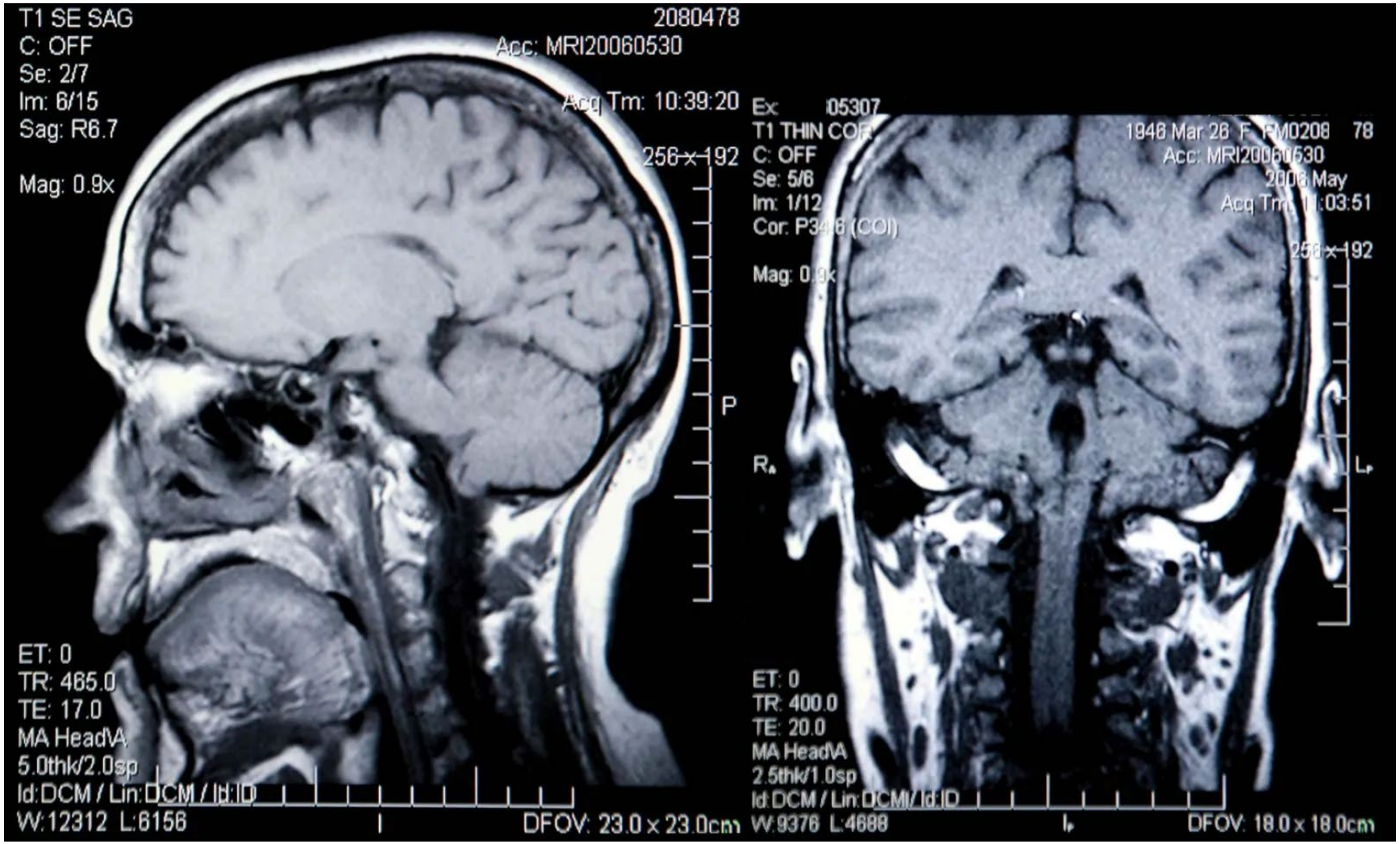
"The de-identification standard makes no distinction between data entered into standardized fields and information entered as free text (i.e., structured and unstructured text) -- an identifier listed in the Safe Harbor standard must be removed regardless of its location in a record if it is recognizable as an identifier." Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule

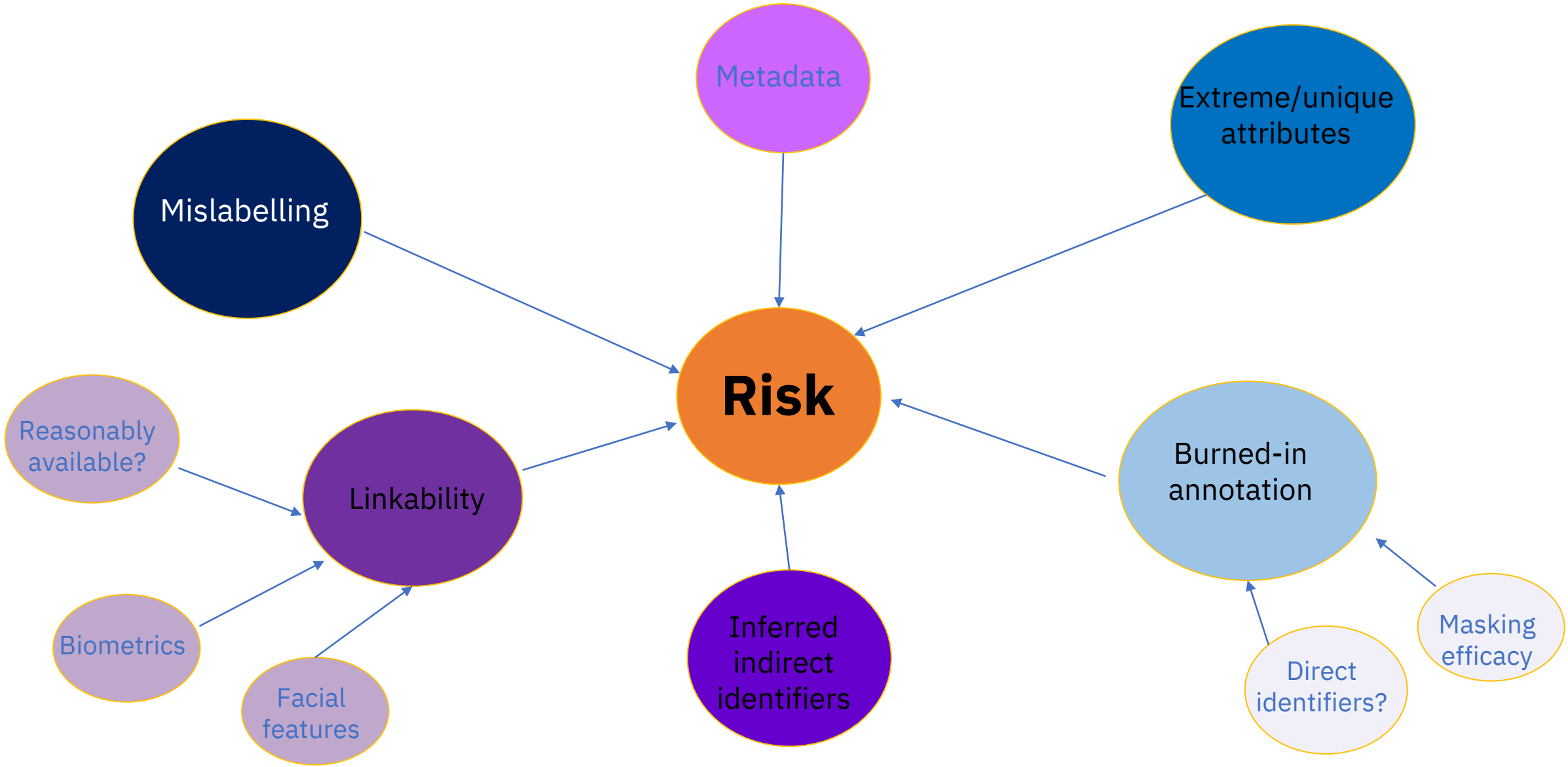
Free-text: Paths to de-identification

- **Named Entity Recognition (NER)** tools can be trained to remove *close* to 100% of identifiers
  - Safe Harbor cannot reliably be achieved at scale by either human annotators or Large Language Models.
  - However, the 'very small' risk standard of Expert Determination may be satisfied by state-of-the-art models.

- **Obfuscation (hide-in-plain-sight) + NER** provides greatest protection.
  - Re-identification depends on *accurate knowledge*
  - Substitution of detected identifiers with plausible synthetic alternatives reduces the recipient's confidence that that any of the few undetected identifiers are in fact real

Medical Images




Metadata
Extreme/unique attributes
Mislabelling
Risk
Reasonably available?
Linkability
Burned-in annotation
Biometrics
Facial features
Inferred indirect identifiers
Masking efficacy
Direct identifiers?

Common modifications required

Privacy+ Security Forum

Image location

Images not present in other datasets

Correction of mismatches between pixel data and metadata

Facial imagery

Either
Redact head fully
or
Apply effective skull stripping

Remove soft-tissue facial pixels (2D X-ray images only)

Metadata

Apply expert's standard restrictions on indirect identifiers

E.g. capping age, removing birth/death date granularity

Limiting geography to state and 3-digit ZIP

Annotation redaction

Ensure masking method captures all text

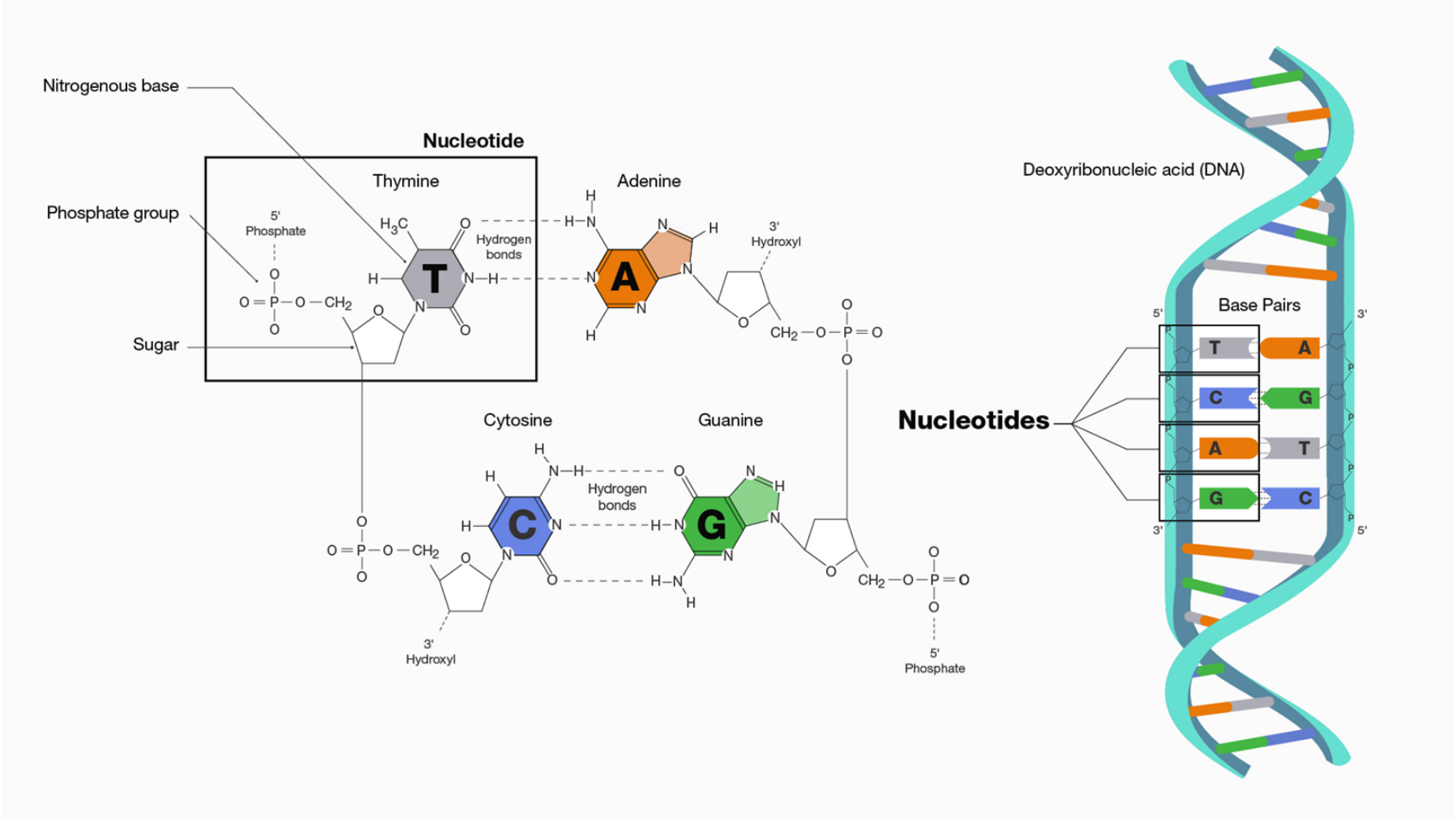
Ensure masking method removes direct identifiers

Remove light bleed of text outside mask

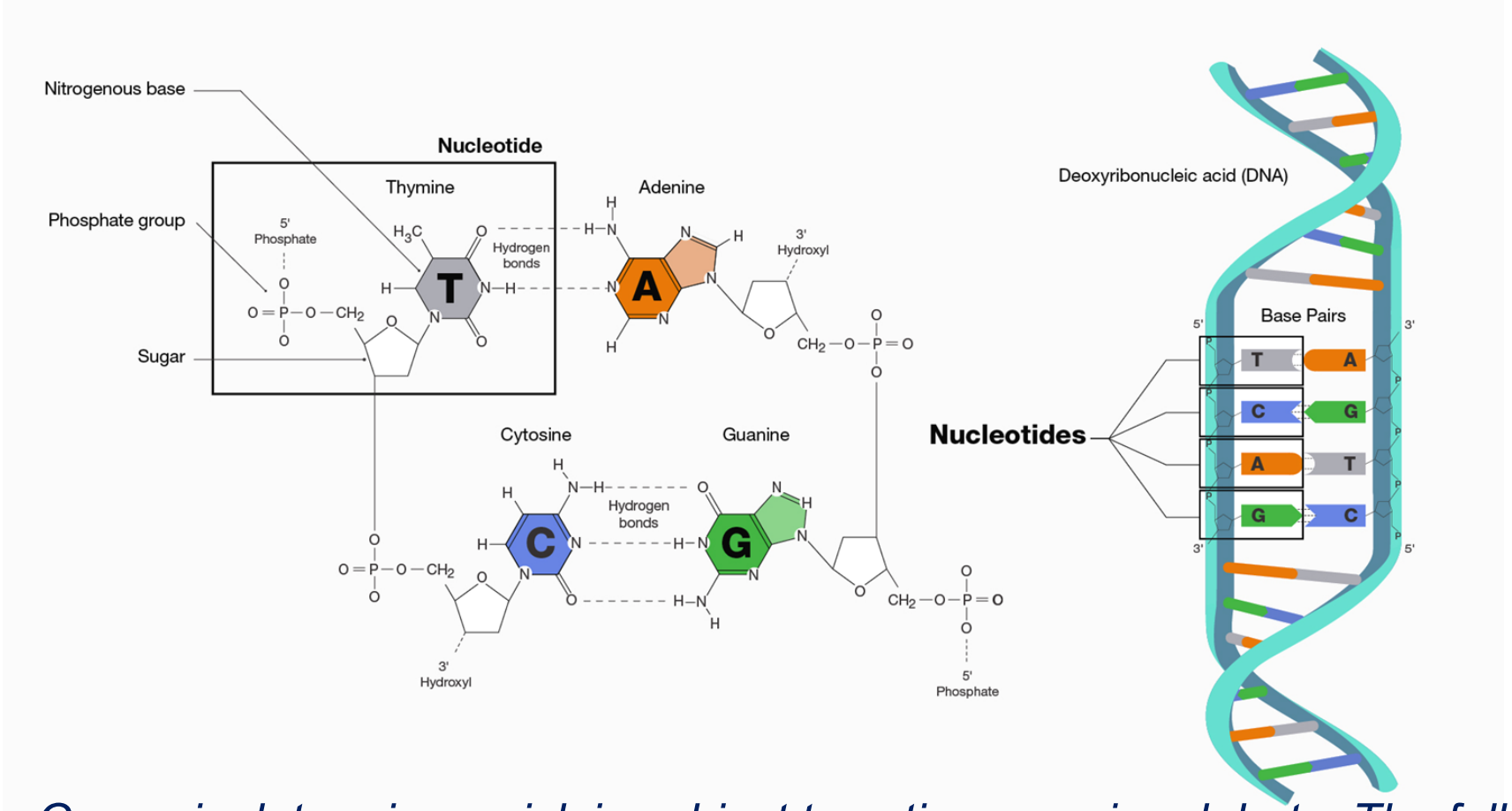
Facial features

- Safe Harbor requires removal of full-face photographs and comparable images
- Identifiable facial features can be reconstructed from volumetric images of head and face structures
- “Skull stripping” - removes all non-brain tissue (most reliable - only axial not sagittal or coronal slices)
- Current de-facing and facial blurring methods do not fully prevent reidentification (Schwarz et al., 2021; Abramian and Eklund, 2019)
- Soft tissue removal possible for X-rays i.e. by setting threshold intensity

Genomic Data



Genomic Data



Disclaimer: Genomic data privacy risk is subject to active, ongoing debate. The following opinions are my own but others may disagree!

Where is the risk? Erlich study

- Most prominent case study, Gymrek et al (2013) led by Yaniv Erlich:
  - Short tandem repeats on Y-chromosome were profiled and combined with querying publicly accessible online genealogy databases
  - Surnames were revealed in an ostensibly ‘anonymized’ dataset for ~135,000 or 12% of individuals
  - Further potential exposure of millions of relations
  - Catalyzed a sea change in public opinion and led to stricter data policies from GWAS, NIH etc.

Legal landscape

- HIPAA Omnibus Rule (2013) classified uniquely identifiable genomics data as PHI (Section 105) following initial codification as health information by GINA
    - Does not include genetics information among the 18 identifiers that Safe Harbor requires for redaction
    - Releasing genomic data under Safe Harbor would present unacceptable disclosure risk and starkly highlights the limited efficacy of Safe Harbor
- The Common Rule *does* classify genomic data as 'non-identified biospecimen'
    - Non-identified ≠ de-identified
    - Not comparable to Expert Determination - CR enables higher risk tolerance for specific, IRB-approved proposals only

Where is the risk? Availability

- Generally risk comes from rare combinations of gene and mutation signifiers (either names or sequence data)

- These carry potential risk of linkage to sources of genetic data including whole exome databases (e.g., the 1,000 Genomes Project), genealogy databases (e.g., Ancestry.com and GEDmatch), published genetic research datasets; and other proprietary genetic information available to the anticipated recipient.

Where is the risk? Somatic vs Germline variants

- Oncology data: **somatic** variants are **safe** (not stable across data source and time so don't facilitate linkage)

- **Germline** variants (heritable; persist across source and time) ***may*** be risky
  - Depends on rarity
  - Depends on combination with other indirect identifiers (age, gender, race etc.)

- How germline variants quickly become **very high risk**:
  - Rarity - e.g. rare disease, only shared by a few individuals in the US
  - Many germline variants per patient - the risk stacks.

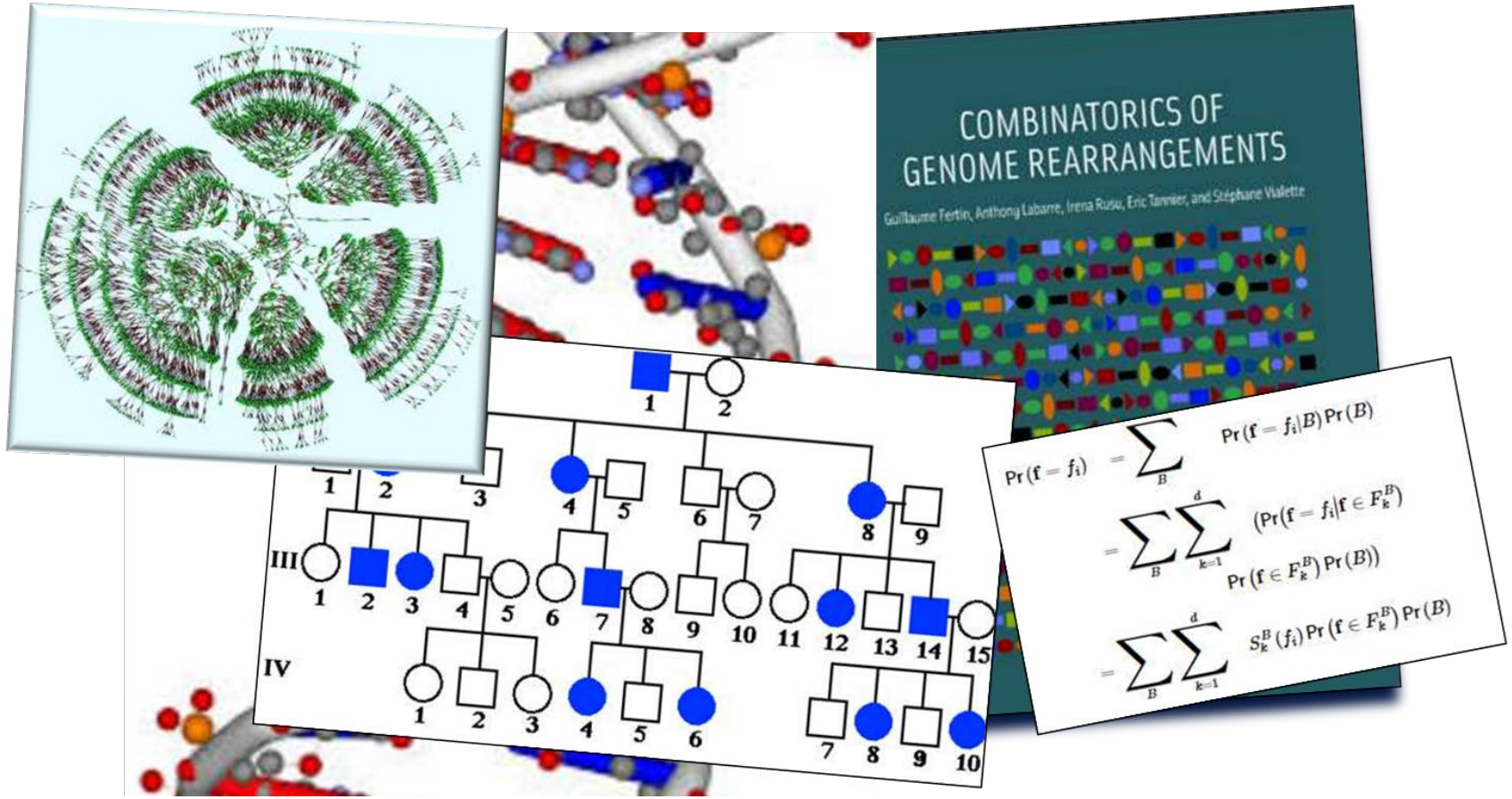
Where is the risk? **Raw Sequence Data**

- BAM/VCF files contain significant proportion of whole genomes
- Not structured data, must be queried
- Essentially unique identifiers - thousands of germline variants in combination
- Maximally high-risk

Where is the risk?

- Genetic diagnostic test and panel names + details
- Biomarker names and +/-ve status
- Categorical and numerical lab data from genetic testing
- Gene name
- Mutation/variant name
- Short nucleotide sequences
- Long sequences
- Raw genomics sequence files e.g. BAM, VCF formats

Genomic combinatorics - family members may also be identified so individual consent cannot satisfy privacy and ethical standards



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

39

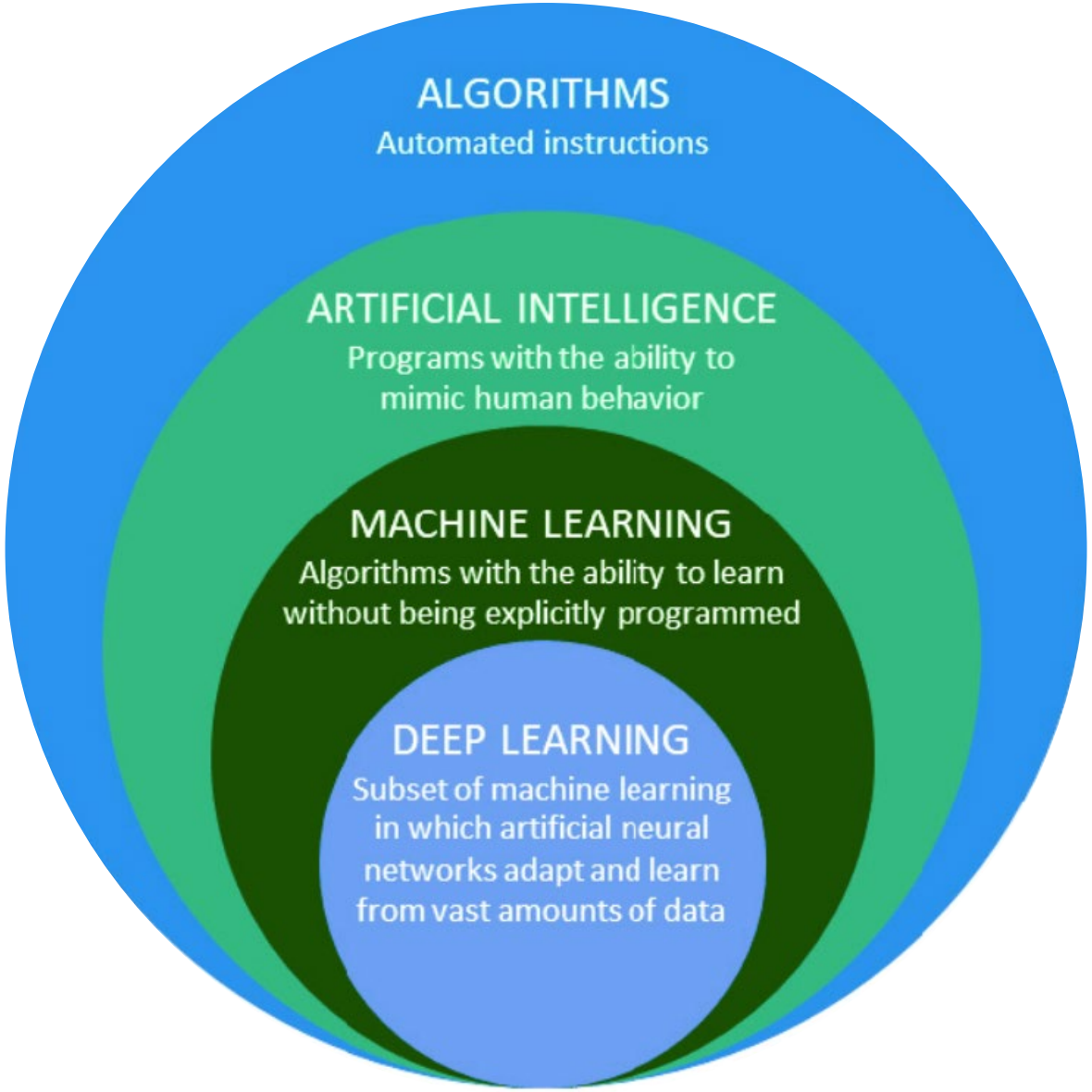
THE CONVERSATION
Academic rigor, journalistic flair

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)

Artificial Intelligence (AI)


ALGORITHMS
Automated instructions
ARTIFICIAL INTELLIGENCE
Programs with the ability to mimic human behavior
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

- Machine Learning models are a game changer for clinical research – identifying disease status, predicting progression with remarkable accuracy

- And rapidly becoming better at classifying individual identifiers – dangers and opportunities for privacy preservation

- *AI Privacy Research is active, exciting and tentative – as a community we do not yet fully understand the scale of risk nor have fully unlocked the benefits of AI*

AI – now and future threats

- Unless *very carefully managed* Machine Learning models trained on one ostensibly de-identified patient cohort risk revealing patient-level information when applied to another patient dataset:

- E.g. accurately predicting patient ZIP code thereby elevating disclosure risk

- For Unstructured Data a human is no longer required to painstakingly sift identifiers from huge volumes of information – a trained model can accomplish it at the touch of a button

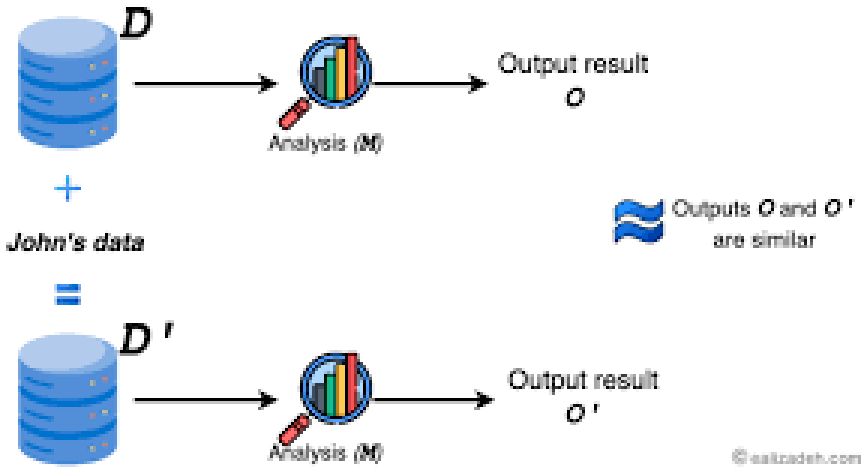
AI – now and future threats

De-facto linkage risk

Exploiting latent identifying trends human experts miss

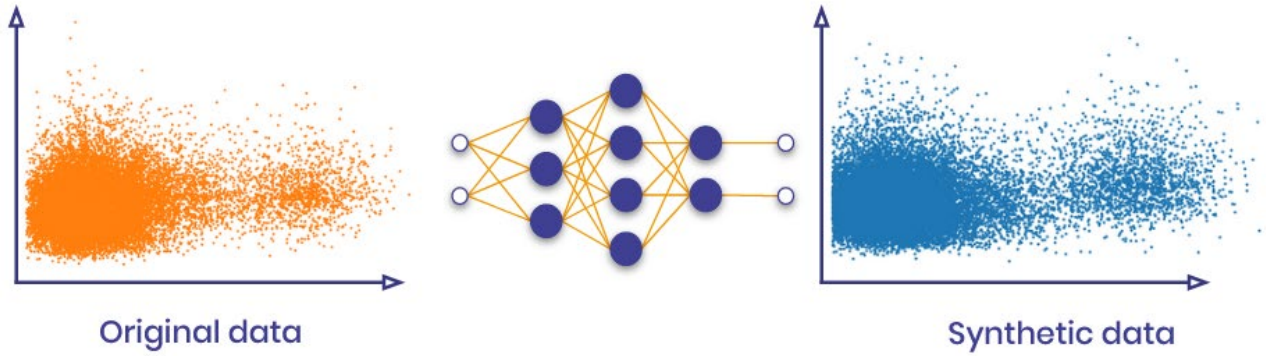
Scalable compute power

AI – nuanced realities



Differential Privacy

Synthetic Data



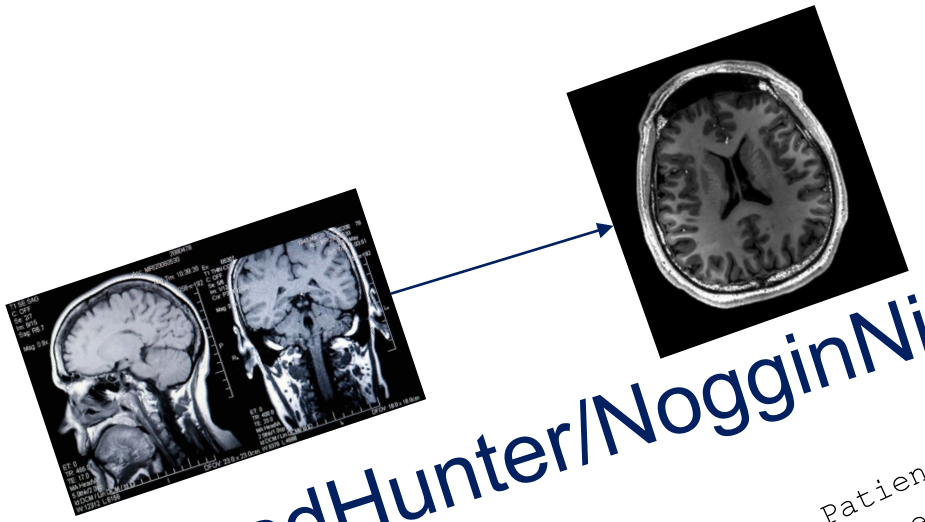
AI – nuanced realities

AI is only as viable as the data it is trained on

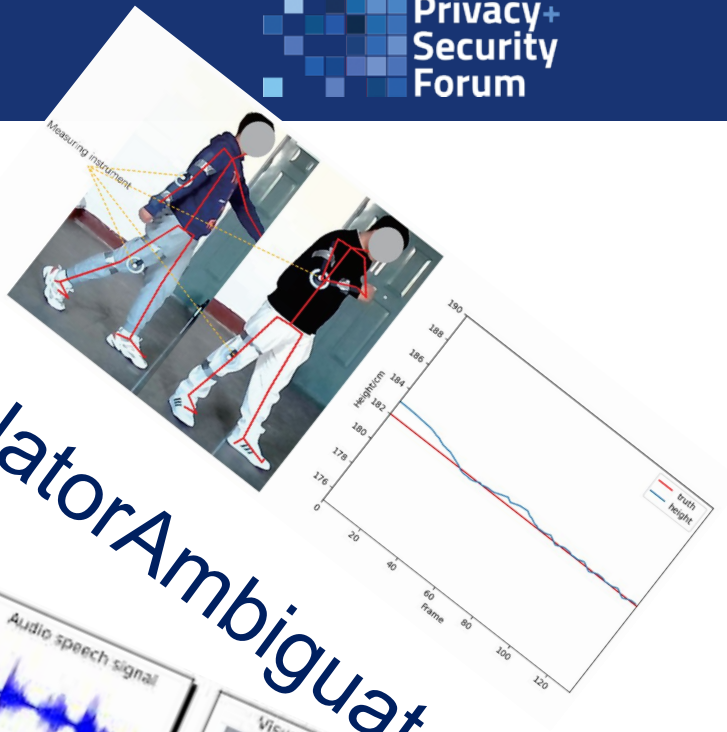
Re-identifiability relies on statistical fidelity → availability of data (images, audio, gait…)

AI technology, data availability and threat models evolve rapidly – risk today will be very different from risk tomorrow

AI – tools for the privacy expert



HeadHunter/NogginNinja

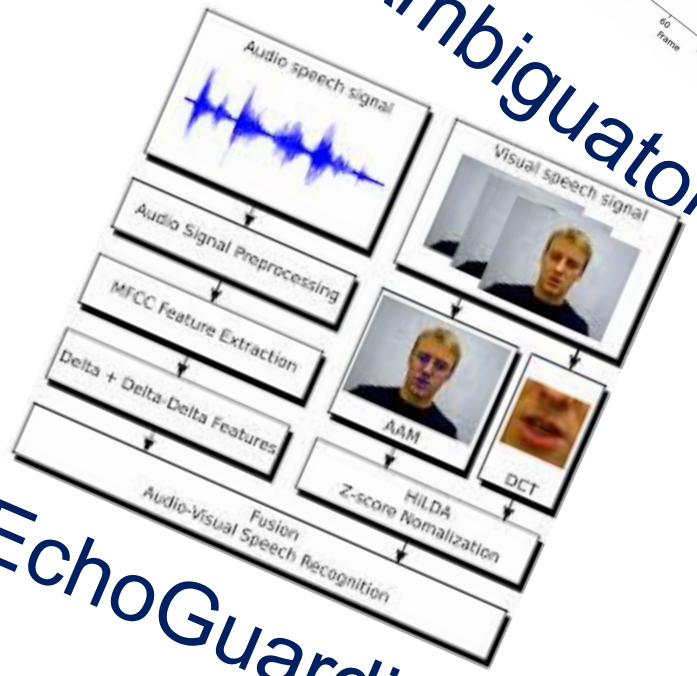


AmbulatorAmbiguator

```
PatientID, Notes, Notes
Surgery_a4, "      Stinson
name:  Ben
PT ID: 6843115 65
DOB: 10/05/1940
```

```
PatientID, Notes, Notes
Surgery_a4, "
name:  NAME
PT ID: PT ID
DOB: DOB
```

BlotOut Bot



EchoGuardian

Break

Privacy+ Security Forum

DE-IDENTIFICATION AND THE LAW(S)

Privacy+ Security Forum

De-Identification Under the New State Laws (And APRA)

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**

  →→→Calendars – Roman, Mayan, Egyptian, Islamic, Hebrew, Hindu, Persian… Gregorian calendar introduced in 1582, finally widely adopted by 19th century, now **the** international calendard standard used WW

  →→→Distance – Scottish mile longer than English mile – Scottish mile outlawed three times!

- **Strong historical trend toward greater harmonization and standardization**

But 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
- Would have resulted in expensive lawyering, contractual wrangling over risk, delays, costs, litigation risk, etc. and generally impeding data research and fluidity

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

De-ID'n under CCPA 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??

18 of the 19 enacted to date (i.e., all except Delaware) have a two-tier structure similar to CA's:

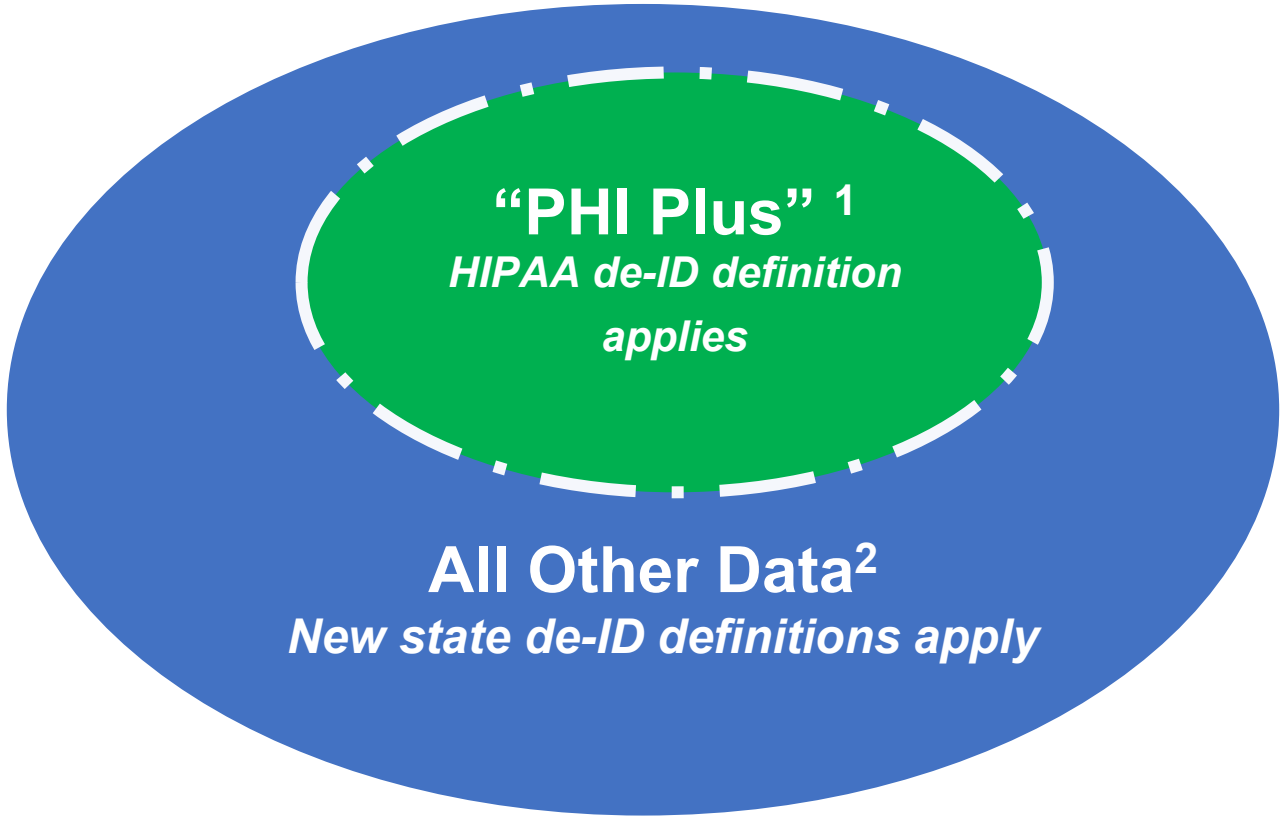
- **HIPAA de-ID'n applies to "PHI Plus" (PHI plus other medical data)**
- **New state-specific de-ID definition applies to all other data**

*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 18 of the 19 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.

[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.



More complexities with de-ID'n under the 18 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).

Privacy+ Security Forum

But wait….

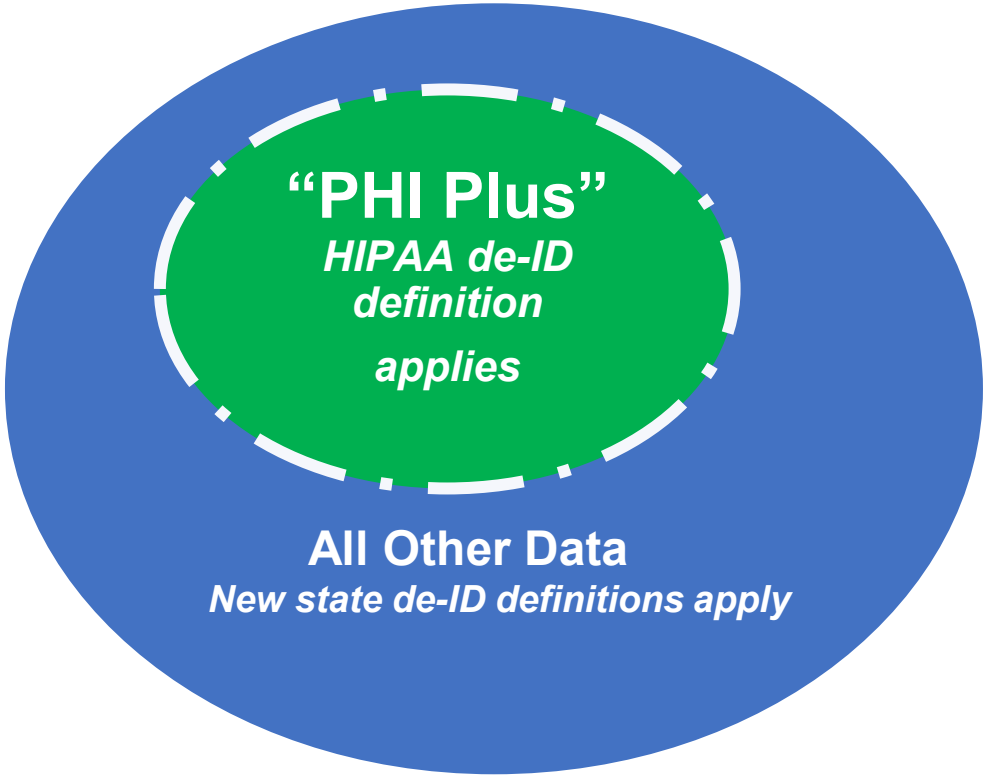
What about Delaware??

Delaware's privacy law:

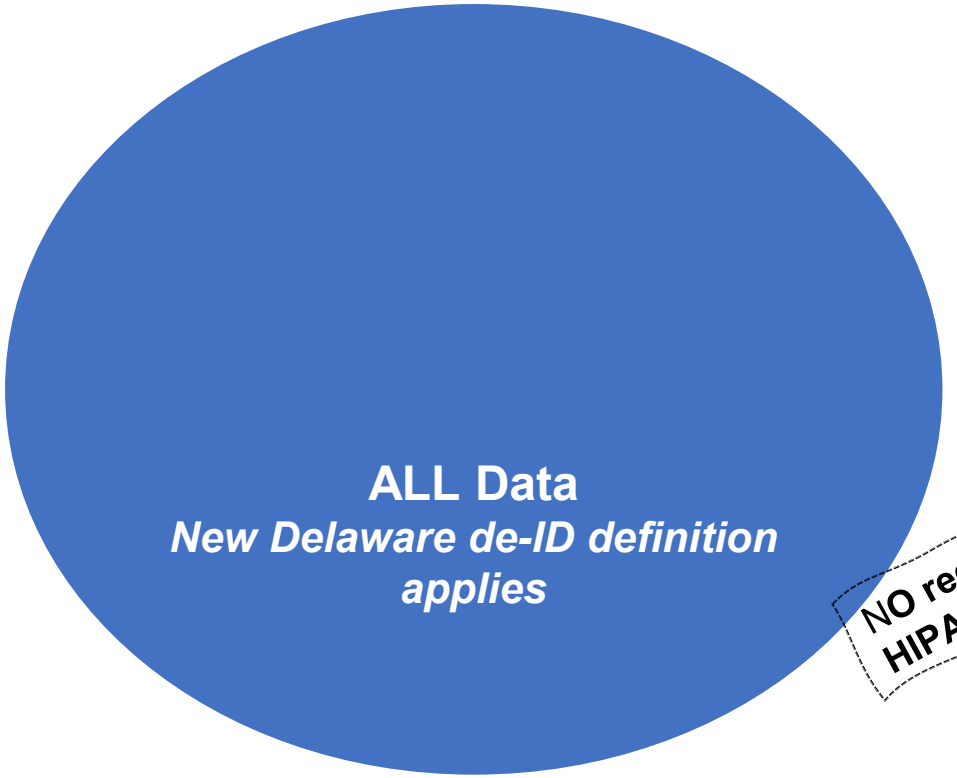
- **Is the ONLY state that does not recognize the HIPAA de-ID standard - not even for PHI**
- Does NOT have a two-tier de-ID structure similar to CA's
- Delaware's general de-ID definition applies to ALL data. It's like the original CCPA (modified in 2020 to harmonize de-ID with HIPAA for "PHI Plus")

Privacy+ Security Forum

18 of 19 State Privacy Laws



Delaware Privacy Law



NO recognition of HIPAA de-ID in DE law

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*

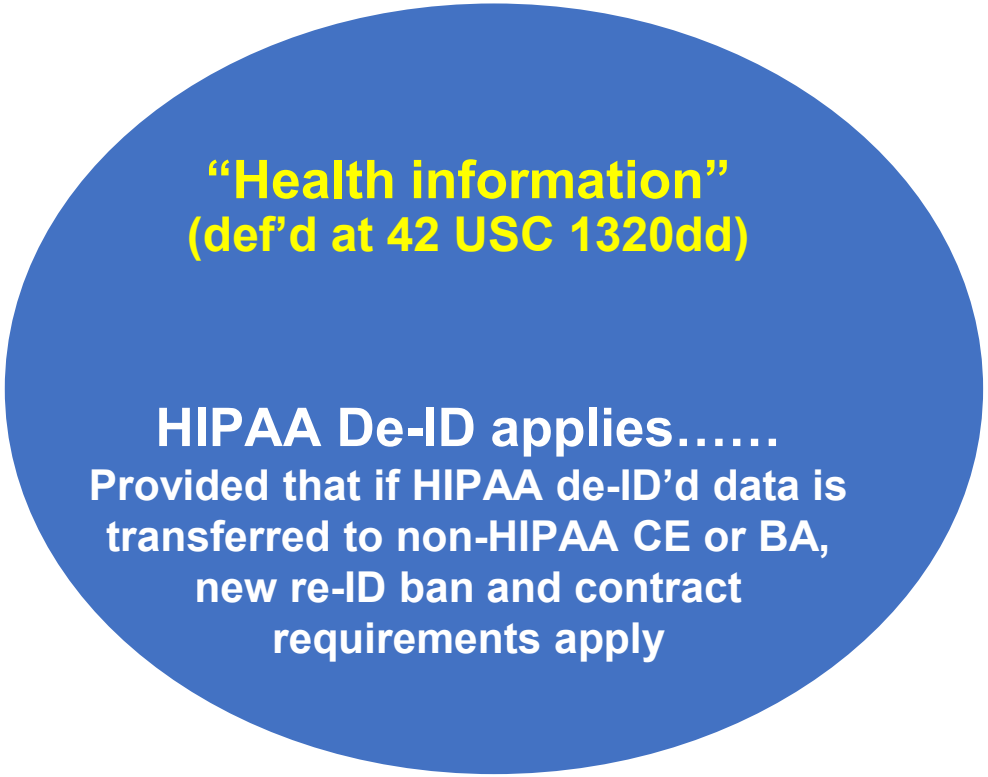
Other New State Provisions Regarding De-ID'n

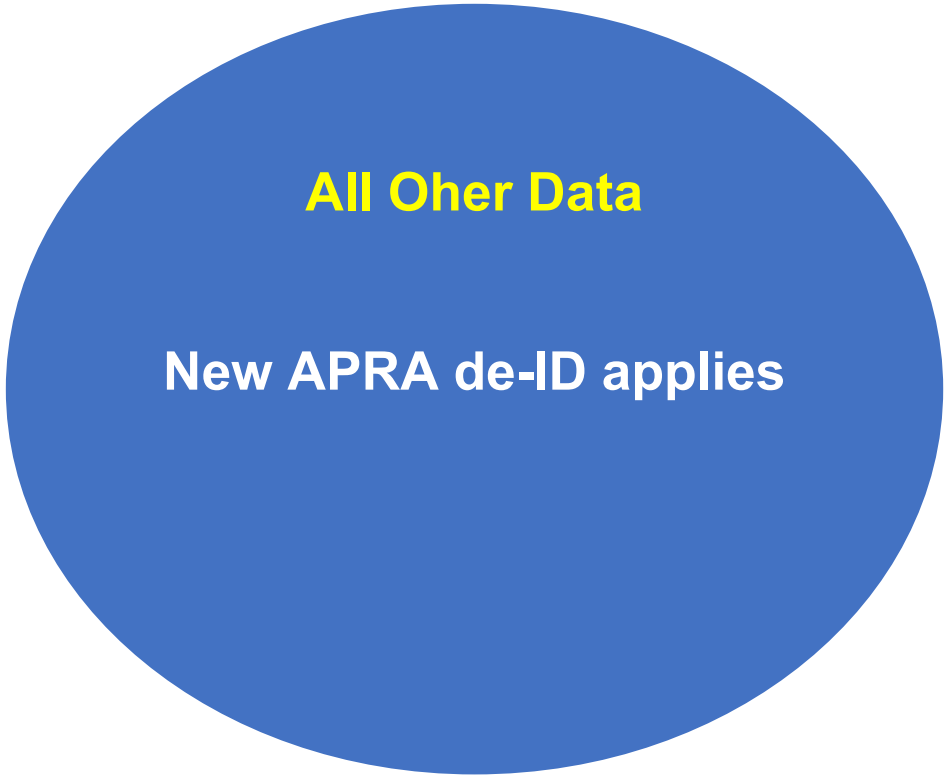
7) Multiple States – Benefits of De-ID'd Data

- 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

De-Identification under the draft American Privacy Rights Act (APRA)

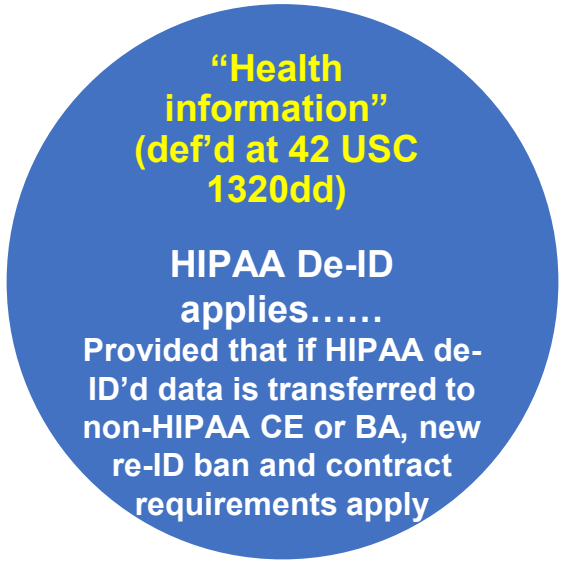
- **APRA follows general state pattern of TWO types of de-ID for different data**
  → *But in novel ways*





New Delaware de-ID definition applies

De-Identification under the draft American Privacy Rights Act (APRA)



SCOPE – Note carefully where the HIPAA de-ID standard would apply. The scope is NOT PHI, nor is it the "PHI Plus" of most state laws. Both broader and narrower –

- **Includes information . . . Created or received by a provider, health plan, public health authority, employer, life insurer, school or university, or HC clearinghouse**
- **Does NOT necessarily include medical research data covered by FDA or Common Rule**

Privacy+ Security Forum

New APRA De-Identification Definition for Non-Health Data

All Oher Data

New APRA de-ID applies

A) information that cannot reasonably be used to infer or derive the identity of an individual, does not identify and

B) is not linked or reasonably linkable to an individual or a device that identifies or is linked or reasonably linkable to such individual, regardless of whether the information is aggregated, provided that the covered entity or service provider—

(i) takes reasonable physical, administrative, or 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;

(ii) publicly commits in a clear and conspicuous manner to—

(I) process, retain, or transfer the information solely in a de-identified form without any reasonable means for re-identification; and

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

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

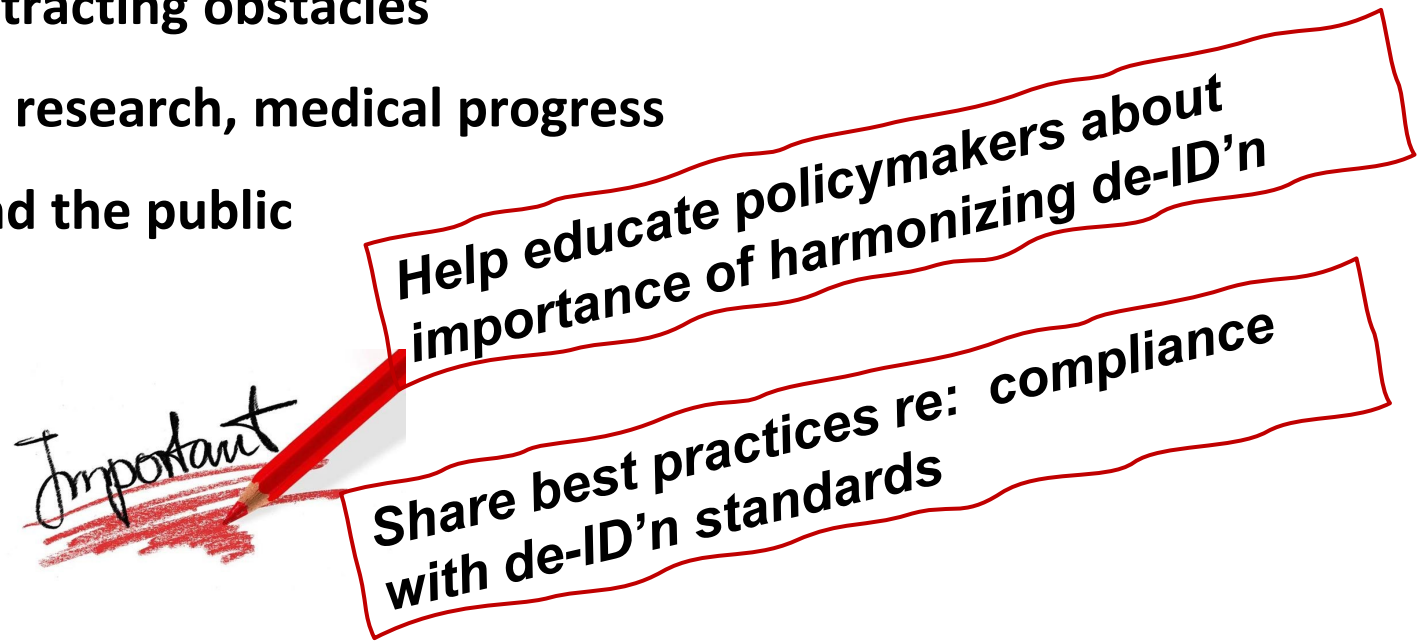
(I) comply with all of the provisions of this paragraph with respect to the information; and

(II) require that such contractual obligations be included in all subsequent instances for which the data may be received;

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



Privacy+ Security Forum

Anonymization – Overview of EU Issues

Legal Context, Practical Implications, Developing Issues

Chris Diaz

Agenda

Privacy+ Security Forum

1. Brief Legal Context & Takeaways
2. Practical Implications
3. Developing Issues

Anonymization Across Globe

- LGPD Article 5 (Brazil)
  - Data relating to data subject who cannot be identified, considering the use of reasonable and available technical means at the time of processing
- LGPD Article 12 (Brazil)
  - Anonymized data shall not be considered personal data, except when the anonymization process to which it was submitted is reversed, using its own means, or when, with reasonable efforts, it may be reversed.
  - What is reasonable must take into account objective factors, such as cost and time require to reverse the process, according to available technologies, and the exclusive use of own resources.
  - May still be personal data if used to form the behavioral profile of a particular natural person, if possible to identify.
  - National authority may dispose of standards and techniques use in anonymization processes and carry out checks on their security.

Anonymization Across Globe

- PIPEDA (Canada)
  - Personal information = “information about an identifiable information”
  - Proposed Bill C-27s explicit definition of “anonymization” = to irreversibly and permanently modify personal information, in accordance with generally accepted best practices, to ensure that no individual can be identified from the information, directly or indirectly, by any means.
- Law 25 (Quebec) – Draft Regulation
  - Information is anonymized if it is, at all times, reasonably foreseeable in the circumstances that it irreversibly no longer allows the person to be identified directly or indirectly…must be anonymized according generally accepted best practices and according to the criteria and terms determined by regulation.
  - Life cycle management of anonymization process
    - Pre-anonymization
    - Anonymization process
    - Anonymization results

Anonymization in GDPR

- Recital 26
  - “Not Applicable to Anonymous Data”
  - 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.
  - [GDPR] does not therefore concern the processing of such anonymous information, including for statistical or research purpose.

Pseudonymization in GDPR

- Recital 26
  - Pseudonymized data that could be attributed to a natural person using “additional information” should be considered information on an ID’d natural person, i.e., personal data.
  - Account should be taken of **all the means reasonably likely to be used** (either by controller or another) to ID the natural person directly or indirectly
  - Examine objective factors such as costs, time required for ID’ing, available technology at time of processing and technological developments.

- GDPR Art. 4(5)
  - Pseudonymization means processing personal data so it can “no longer be attributable to a specific data subject without the use of additional information, providing such info is kept separate and subject to [TOMs] to ensure that the personal data are not attributed to an [ID’d] natural person.”

Privacy+ Security Forum

Art. 29 Working Party Opinion (2007)

- WP Opinion 04/2007 – Section III.3. – "Identified or Identifiable" [Natural Person]
- This means "**mere hypothetical possibility to single out the individual is not enough to consider the person as 'identifiable**'."
- Where natural person cannot be ID'd whether by data controller or other person, taking int account all the means likely reasonably to be used to ID that individual. (pg. 21)
    - Assessment of whether the data allow identification of an individual, and whether the information can be considered as anonymous or not **depends on the circumstances**, and a **case-by-case analysis** should be carried out with particular reference to the extent that **the means are likely reasonably to be used** as described in Recital 26.
    - This is particularly relevant in cases of statistical information (e.g., aggregated data where original sample not sufficiently large and other pieces of information may enable identification)

Art. 29 Working Party Opinion (2014)

- WP Opinion 05/2014 – Section 2 – Definitions and Legal Analysis
  - Technique ”applied to personal data in order to achieve **irreversible de-identification**. (pg.. 7)
  - Anonymisation process” is “further processing” that must comply with test of compatibility…” (pg. 7)
  - “…pseudonymized data cannot be equated to anonymized information as they continue to allow an individual data subject to be singled out and **linkable across different data sets**.” (pg. 10)
  - “Thus, it is critical to understand that **when a data controller does not delete original (identifiable) data** at event-level, and the data controller hands over part of this dataset…**the resulting dataset is still personal 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.” (pg. 9)
  - “An effective anonymization 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.

Guidance Documents

- CNIL Guidance
  - “Anonymization is a treatment which consists of using a set of techniques in such a way as to **make it impossible**, in practice, for any identification of the person by any means whatsoever and **in an irreversible manner**.”
  - https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

- Irish DPC
  - “’Anonymization’ of data means processing it within the aim of **irreversibly preventing** the identification of the individual to whom it relates. Data can be considered effectively and sufficiently anonymized if it does not relate to an identified or identifiable natural person or where **it has been rendered anonymous in such a manner that the data subject is not or no longer identifiable**.”
  - https://www.dataprotection.ie/en/dpc-guidance/anonymisation-and-pseudonymization

Guidance Documents

- EDPS and AEPD Joint Guidance on Hashing
  - "…anonymization procedures must ensure that **not even the data controller is capable of re-identifying the data holders** in an anonymised file."
  - https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

- EDPS FAQs on Anonymization, Misunderstanding #5
  - "Although a 100% anonymization is the most desirable goal…in some cases it is not possible and a **residual risk of re-identification must be considered**."
  - https://www.edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en

Privacy+ Security Forum

“Guidance” Documents (UK)

- UK ICO – Consultation – Chapter 1: introduction to anonymisation
  - ”It is important to note that you must carefully assess each case individually based on the specific circumstance…"
  - “This means that even where you use anonymization techniques, a level of inherent identification risk may still exist. However, this residual risk does not mean that particular technique is ineffective. Nor does it mean that the resulting data is not effectively anonymized for the purposes of data protection law when you consider the context.”
  - Also, data protection law does not require anonymization to be completely risk-free. You must be able to mitigate the risk of re-identification until it is sufficiently remote that the information is ‘effectively anonymized.’”
  - https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/

Relevant Cases

Privacy+ Security Forum

C-582/14 – *Breyer (2016)*

- Issue/First Question:
  - Is a dynamic IP address registered by an online media services provider when a person accesses a website that that provider makes accessible to the public constitutes, with regard to that service provider, personal data within the meaning of that provision, where, only a third party, in the present case the internet service provider, has the additional data necessary to identify him?
- Holding:
  - Yes, the dynamic IP constitutes personal data.
- Rationale:
  - Online media service provider has the means likely reasonably to be used in order to identify the data subject with assistance of other persons in context of criminal proceedings, e.g., during cyber attack. The OMSP has the "legal means" to access additional information to identify the data subject.

Was *Breyer* supportive of the objective/non-flexible view of anonymization, or more supportive of the contextual/risk-based view of anonymization? Could it be seen as both?

Relevant Cases

Privacy+ Security Forum

T-557/20 – *SRB v. EDPS (2023)*

- Issue:
  - Was it appropriate for the EDPS to conclude that the information provided to Deloitte qualified as personal data because the sender, SRB, held additional information that could identify original authors of the information?
- Holding:
  - No, the EDPS did not have authority to deem the information personal data.
- Rationale:
  - EDPS **did not investigate** whether Deloitte had legal means available to which it could in practice enable it to access the additional information necessary to re-identify the authors of the comments. It was for EDPS to determine whether the possibility of combining the information that had been transmitted to Deloitte with the additional information held by the SRB constituted a means likely reasonably to be used by Deloitte to identify the authors of the comments.

How far should we read *SRB*? Is it merely indicating the 'test' needed to be applied or did it also determine that the data set from Deloitte's perspective was not personal data?

Relevant Cases

C-319/22 – *Scania (2023)*

- Issue:
  - Does a VIN constitute personal data?
- Holding:
  - Yes, so far as the person who has access to a VIN may have means which reasonably allow them to use the VIN to identify the owner to which it relates.
- Rationale:
  - Court references Article 4(1) GDPR relating to definition of 'personal data' and paragraphs 42 – 43 of the *Breyer* decision. This was also a position stated by the AG, i.e., opinion being VINs are personal data to independent operators 'where [they] may reasonably have at their disposal the means enabling them to link a VIN to an identified or identifiable natural person."

Privacy+ Security Forum

Legal Context – General Takeaways

- Absolute concept of 'anonymization'
  - No risk of re-identifiability exists, not really assessed from perspective of the entity holding the data set in question
  - Seems to be growing trend that this may not be feasible approach, particularly in light of technological developments
- Risk-based/contextual concept of 'anonymization'
  - A residual risk of identifiability may exist, the question hinges on whether it's at an acceptable level
  - Approach supports assessing the entity's position in relation to the 'anonymized' data set, i.e., from the view of the recipient who may not have all 'additional information' in their possession or disposal
  - Trajectory tipping towards this risk-based/contextual approach, inclusive of a "pseudonymized+"
- There are nuances within 'pseudonymization' and 'anonymization', impacted by sociolegal controls.
  - Some arrangement of 'pseudonymization+' may meet legal requirements for anonymization
  - Technical, organizational measures; data recipient/user credentialing; participation requirements; data use agreements, etc.
- Will still always require case-by-case analysis, involving assessing means likely reasonably to be used test, including examination of "legal means"

Privacy+ Security Forum

Achieving 'Anonymization' – Basic Questions

- Many organizations are looking to pseudonymized or anonymized data sets to mitigate the privacy risks associated with data sets but this raises more questions
  - Does your organization have the internal technical capacity to achieve anonymization? The right talent, infrastructure and technology?
  - Even with the resources, how can organizations meet the legal requirements of anonymization, when there is still ambiguity on what qualifies as anonymization? Will there be attempts at full anonymization or pseudonymization+ (basic or strong w/additional sociolegal controls)? What about ISO 27559?
  - How should an organization measure an acceptable level of residual risk of re-identification? How often does a statistical analysis on the residual risk need to be conducted? Who decides?
  - Is it worth the investment based on the ultimate utility of the resulting data set?
  - Even with a legal basis for processing the original personal data, are their ethical considerations that need to be considered?

Ultimately, it's a risk profile question that's case-by-case, akin to how the various cases have examined the question on anonymization. Context will always impact your answers here.

Achieving "Anonymization" – Additional Layers

- What is the feasibility of creating anonymized synthetic data from source? Will you maintain the utility?
  - Create anonymized data from source data by retaining and reflecting relationship/correlations in the source data but not the original data itself
  - Examples:
    - Medidata's Simulants technology, see U.S. Pat. No. 11,640,446
    - WeData/Octopize avatar technology, audited by CNIL
- What about federated approaches in conjunction with anonymization standards?
  - Training at the location of the sensitive data source. Outputs if trained on non-anonymized could still result in privacy risks, but what if there was anonymization to the sensitive data source beforehand?
  - How would this impact the analysis on access to legal means to other data sources?
  - Does this meet the specific requirements for your use case? For example, how does it impact traceability?

Would such additional technologies layered on top of your anonymization processes support legal determination that the resulting data set is 'anonymized'?

Privacy+ Security Forum

Managing Vendor/Sub-Processor Risks

- There are common issues in dealing with vendors, sub-processors who make claims that they will only use “de-identified” or “anonymized” data for their product improvements, internal business uses
  - Often there is reliance on “best industry practices” relating to anonymization
  - Most of these agreements are negotiated with stakeholders who may not fully understand the nuances in terminology, which opens risks to your organizations
- What are best steps to manage these risks?
  - Establish a framework, or supplement existing one, for vendor due diligence
    - Will you always require expert determination? What is an absolute ‘no’? Do you measure them against a standard, ISO? Is there an ‘anonymization’ RACI, does your data set profile need one?
  - Operationalize
    - Clearly define the terms you need as part of your overall vendor/sub-processor risk playbook
    - Internal training for procurement teams, contract managers, and attorneys
    - Process for dealing with escalation, i.e., legal? or data set owner?
    - What architecture will you need for the appropriate data sharing/access?

Developing Issues

Privacy+ Security Forum

EU Data Strategy

- How will EU Data Strategy impact our definitions of anonymization?
  - Data Governance Act
    - Reliance on member states having the technical means for anonymization, pseudonymization to engage in the sharing framework
  - Data Act
    - Applicable to personal data and “non-personal data”
    - Sharing of personal data that has been ‘anonymized’, see e.g., B2G sections
  - Common European Data Spaces
    - European Health Data Space
      - Secondary use of Electronic Health Record (EHR) data, see e.g., Recitals 43, 50, 64

EDPB Guidelines on Anonymisation

- Listed as part of the EDPB Work Programme 2023/2024

Developing Issues

Privacy+ Security Forum

Artificial Intelligence

- As alluded earlier in the workshop, AI/GenAI will continue to change the nature of re-identification attacks.
- How will these risks be incorporated in a functional, harmonized legal definition of ‘anonymization’?
- What impacts will this have on expert determinations, ability to reduce residual risks to an acceptable level?

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

Chris Diaz

Medidata Solutions, Inc.

chris.diaz@3ds.com

David Copeland, PhD

Senior Data Scientist and Privacy Expert
Privacy Hub by Datavant

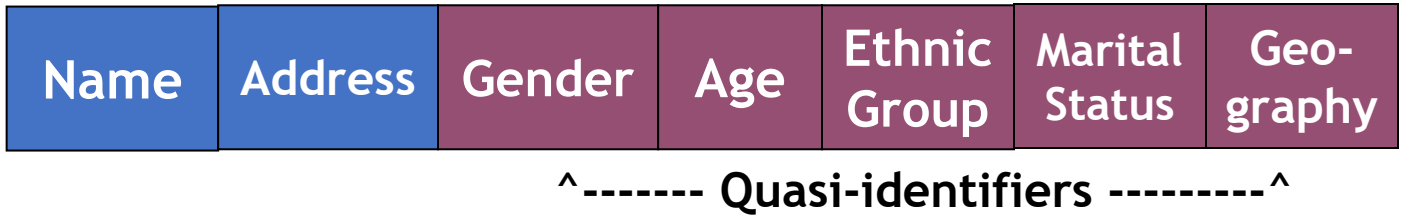
davidc.privacyhub@datavant.com

Privacy+ Security Forum

Reference Slides

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***.



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.).

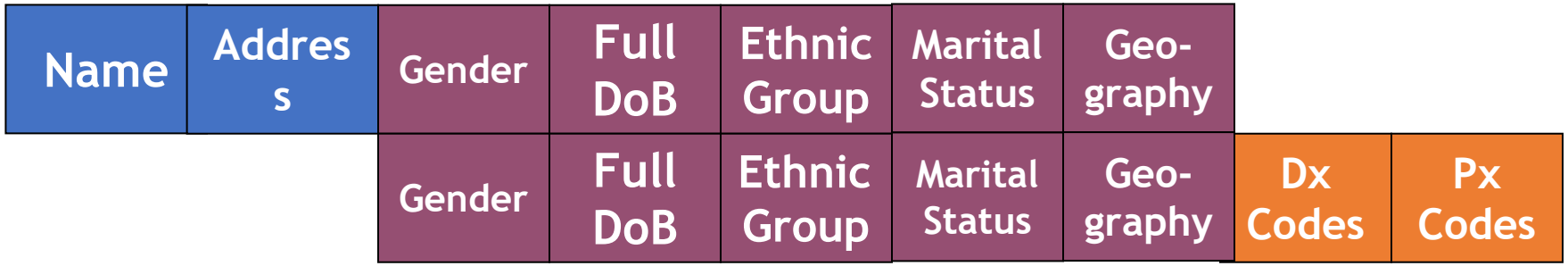
Note that this includes even fields that are not "PHI".

Key Resolution

Key "*resolution*" exponentially increases with:

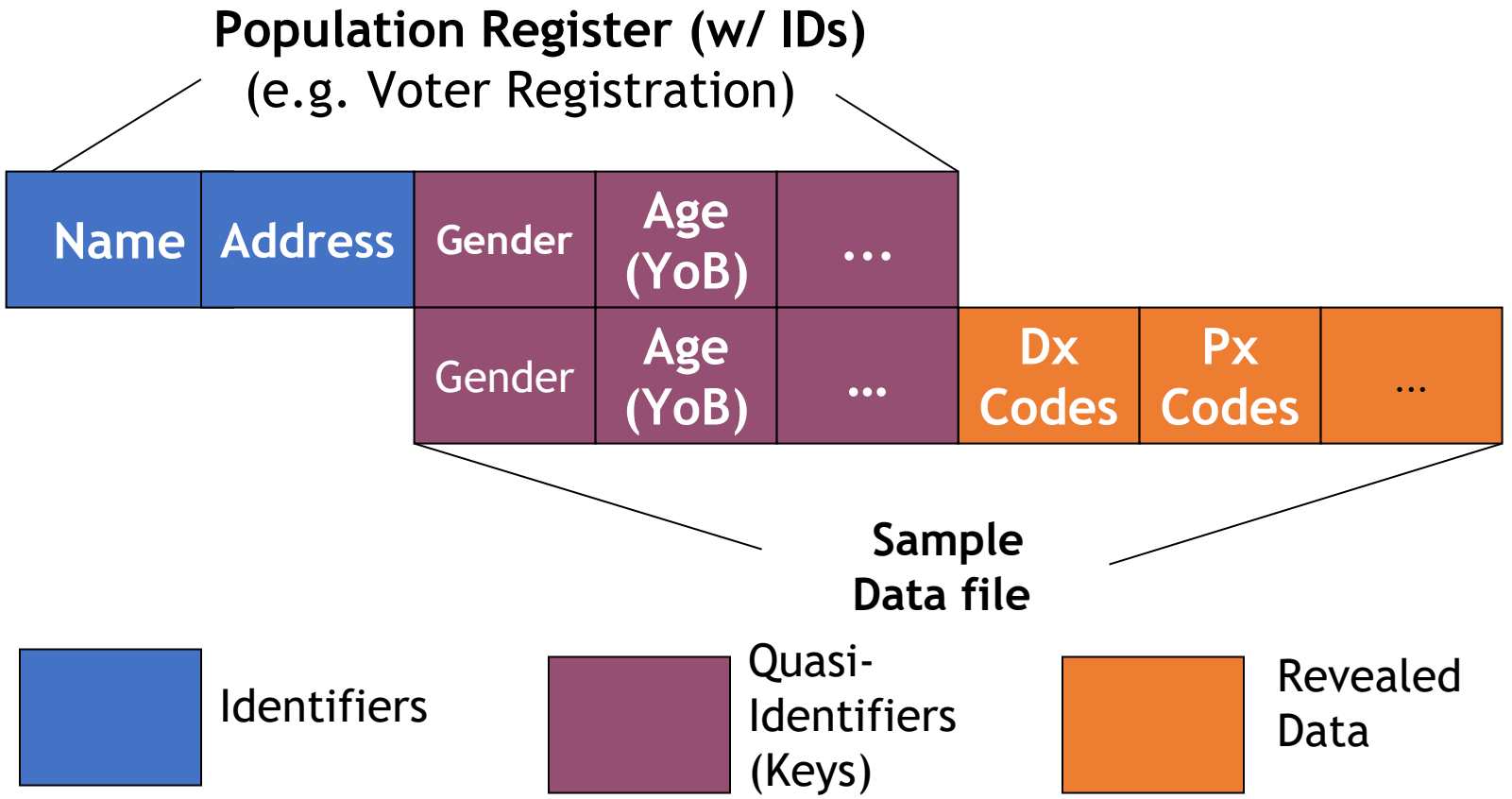
1) the number of matching fields available

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



Record Linkage

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



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).

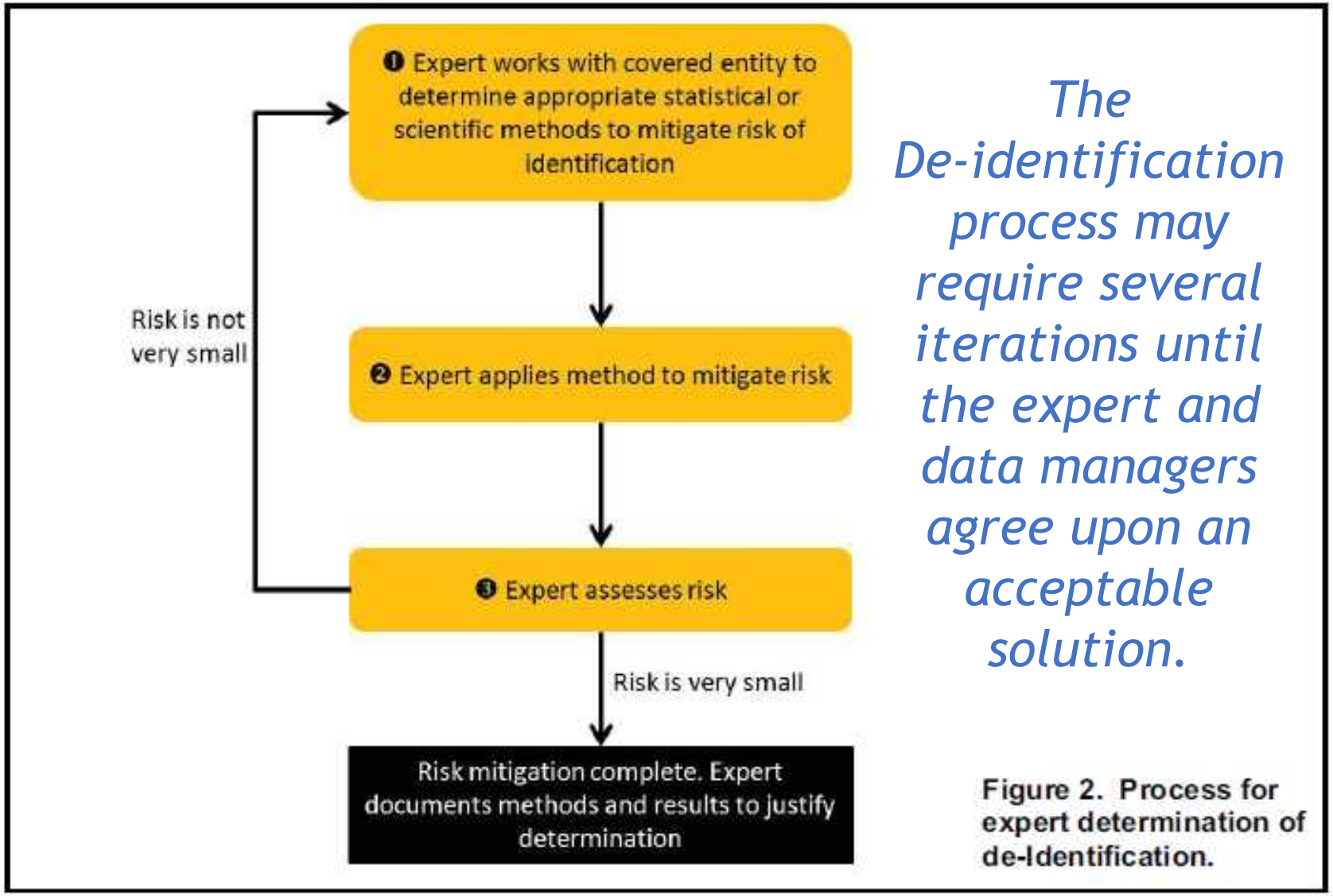
Q2.5 *Can an expert derive multiple solutions from the same data set for a recipient?* *(Cont'd)*

- 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



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.

Audience Question

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

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

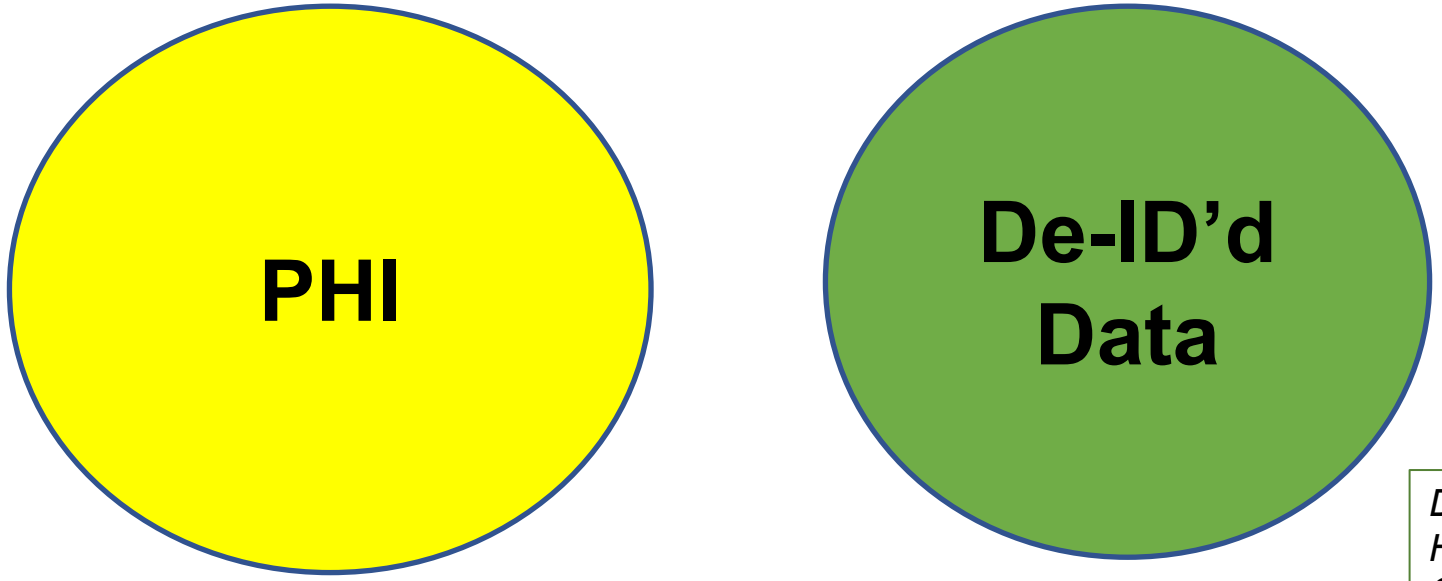
Audience Question

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

Privacy+ Security Forum

De-identification under HIPAA - Basics

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



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