May 10, 2023

# De-identification, Pseudonymization, Anonymization and Cryptographic Tokenization for Privacy Lawyers and Compliance Managers

**Daniel Barth-Jones, PhD**
Principal Privacy Expert,
Privacy Hub by Datavant

**Ann Waldo, JD**
Waldo Law Offices

**Peter Dumont**
Chief Privacy Officer
Optum Labs

**Claire Manneh, MPH**
Head of Provider Research
Datavant

Privacy+ Security Forum

# Speaker

**Daniel Barth-Jones, MPH, PhD**
Principal Privacy Expert,
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

**Claire Manneh**
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

**Ann Waldo**
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

**Peter Dumont**
Chief Privacy Officer,
Optum Labs

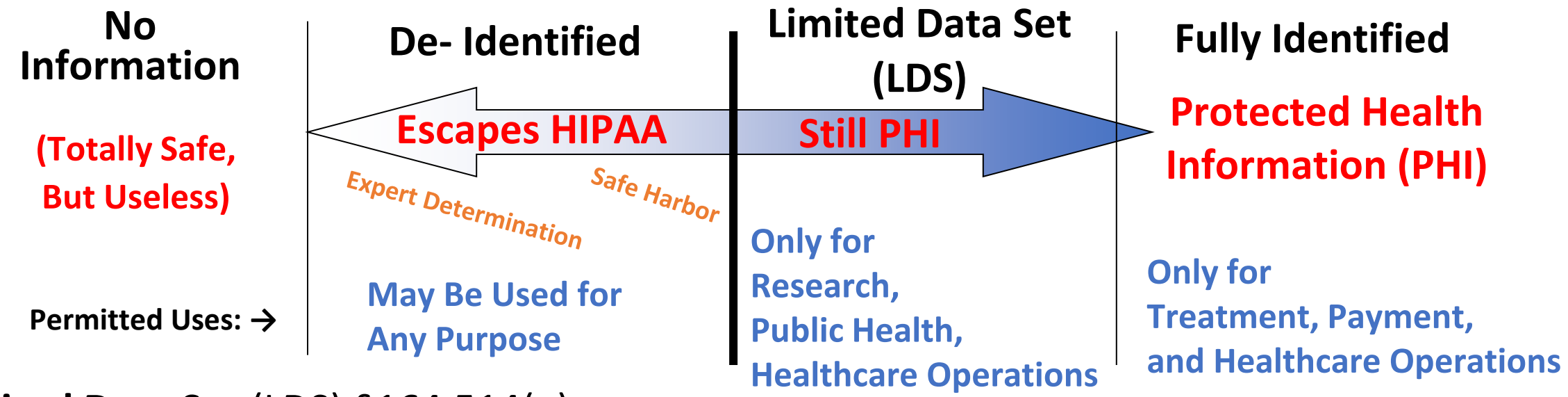
Peter Dumont serves as Vice President and Chief Privacy Officer at Optum Labs. He leads privacy at Optum Labs, and has managed de-identification programs for UnitedHealth Group for over 12 years. At UnitedHealth Group, he has held senior leadership roles in information security, privacy, and data governance. He was a member of the HITRUST De-identification Working Group assisting in developing and issuing the HITRUST De-identification Framework and was Co-chair of the Workgroup for Electronic Data Interchange (WEDI) Security and Privacy Workgroup from 2015 to 2018, supporting their advisory role to HHS. Peter contributed to the National Institute of Standards and Technology Interagency 8053 report on De-Identification of Personal Information. He holds his Certified Information Privacy Professional/United States, Certified Information Systems Security Professional, and HITRUST Certified De-identification Expert certifications. He has a B.S. in Computer Science from the United States Naval Academy and a Masters in Information Technology. Peter served in the U.S. Marine Corps for five years. Prior to joining Optum in 2006, he was an Information Risk Manager with KPMG, LLC.

## 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 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?

# Framing De-Identification under HIPAA and EU Law

# HIPAA's Identification Risk/Legal Spectrum




**Limited Data Set** (LDS) §164.514(e)

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

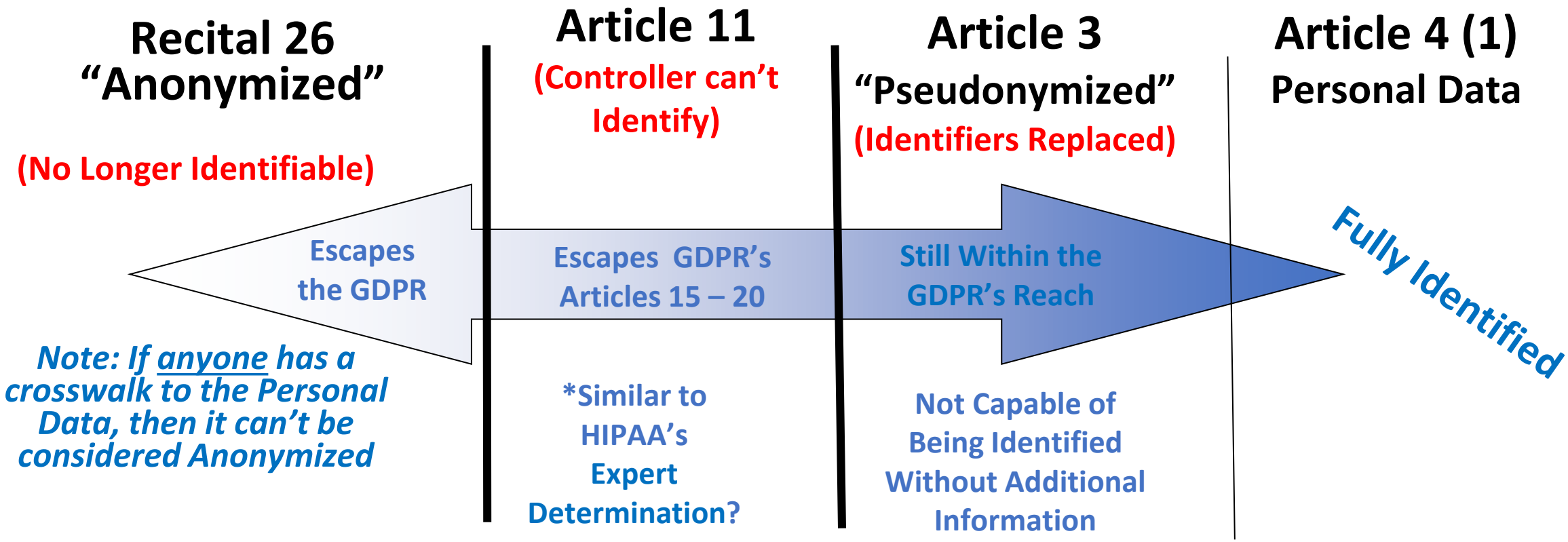
**Safe Harbor De-identified** §164.514(b)(2)

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

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

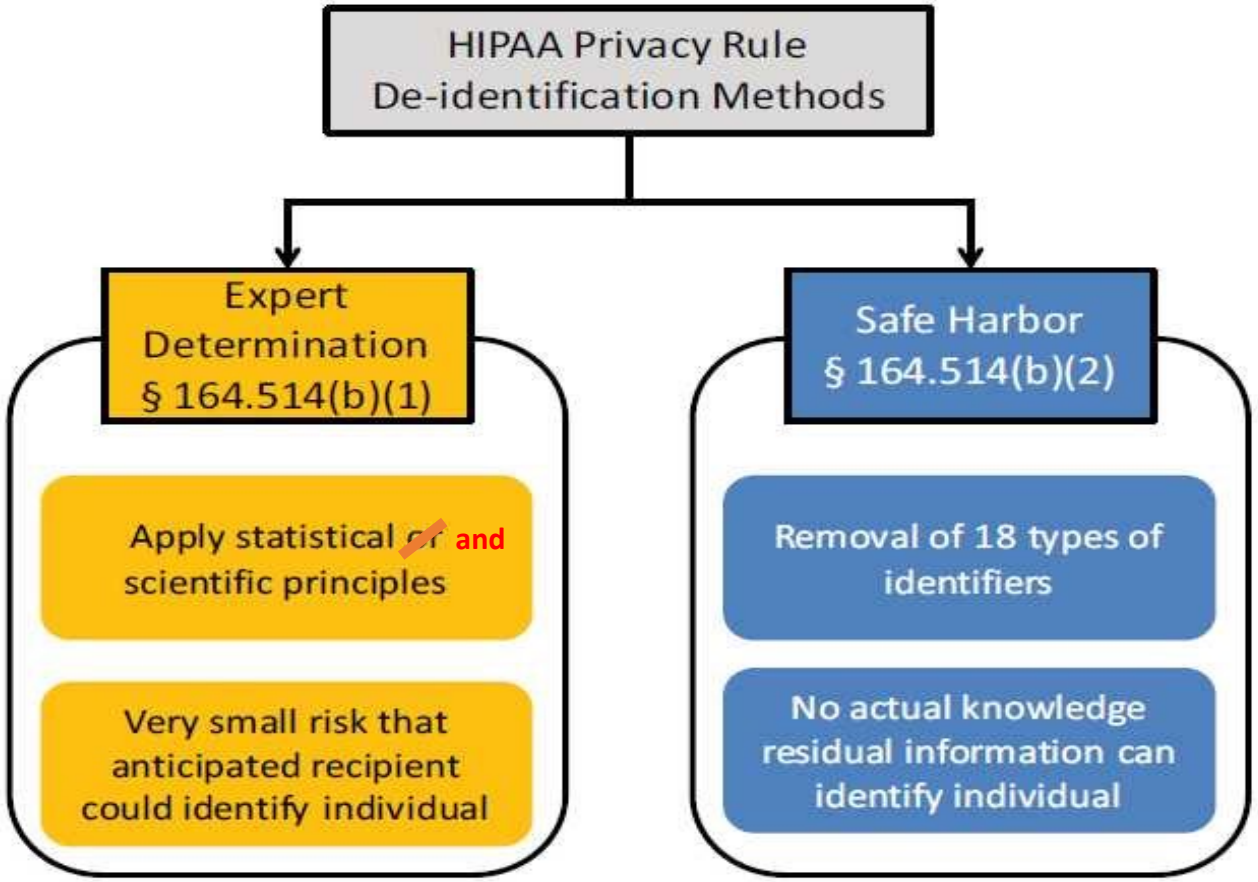
Expert's Analysis Confirms a "Very Small" Risk of Re-identification

# GDPR's Identification Risk/Legal Spectrum




*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

# Two Methods of HIPAA De-identification




Source: 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*:

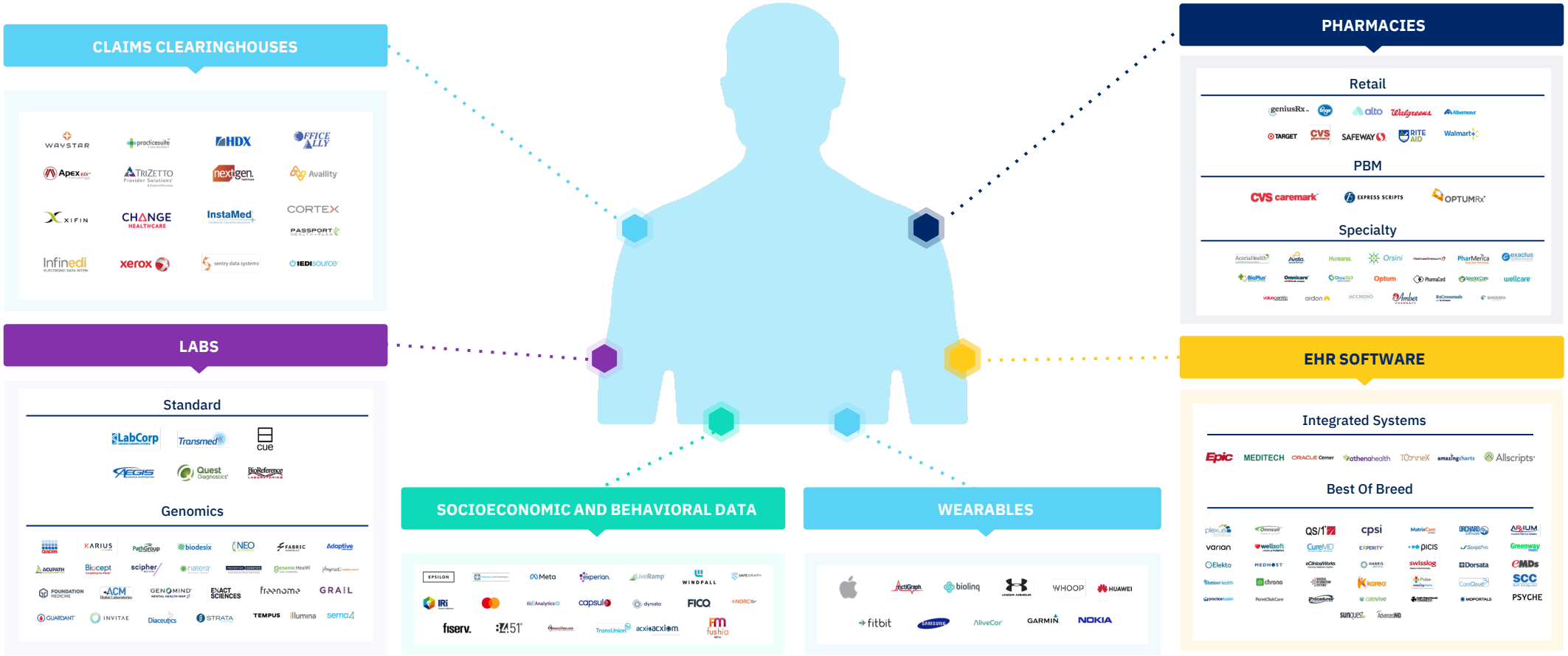
(i) Applying such principles and methods, determines that the *risk is very small* that *the information could be used*, alone or *in combination with other reasonably available information, by an anticipated recipient to identify an individual* who is a subject of the information; and (ii) Documents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

## Questions

- What are pros and cons of Safe Harbor vs. Expert Determination methods of HIPAA de-identification?
- What to consider when deciding between a Limited Data Set (LDS) vs De-identified data?
- What are the unique concerns or issues to consider when integrating consumer or commercial data into De-identified data?
- What to consider when evaluating data retention issues with De-identified data?
- What are ways to consider employing De-identified data to reduce privacy risk?

# How are Researchers Working with Health Data Today?

# Accurate decisions require the ability to connect patient data, no matter where it lives



# 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

# There were pockets of connected patient data, but the industry lacked a standard

# 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*

**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

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.

# Records Linkage

- HIPAA prohibits the sharing of identifiable individual health information 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 suppliers
- 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 is

# Tokenization: A potential solution

# Privacy-Preserving Record Linkage (PPRL)


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

**Patient 12**

**Jane Doe**
**DOB: 05/14/1968**
**SSN: 362-52-1066**
**Phone: 660-945-3120**

Arrived at pharmacy on X date
to pick up DEF prescription

**Patient 34**

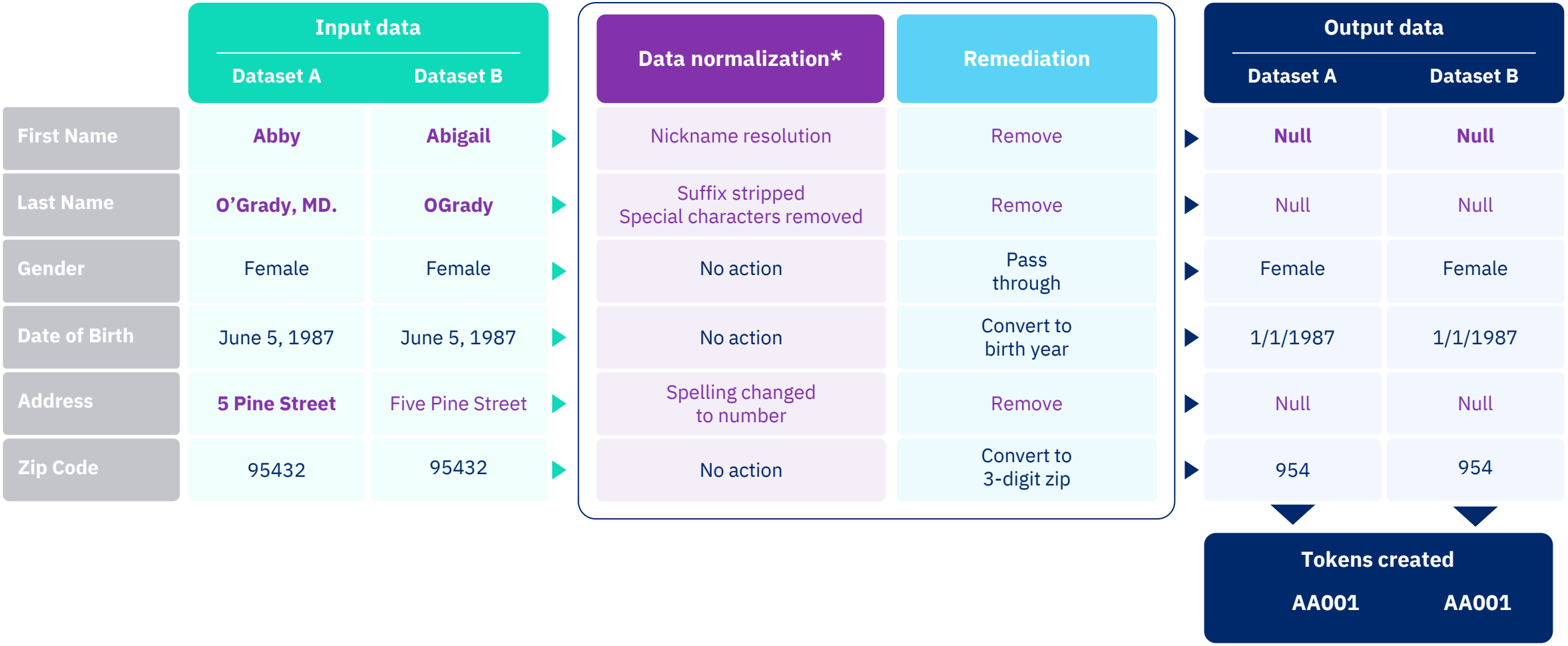
**Patient Key:**
**Patient 12**

**Patient Key:**
**Patient 34**

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

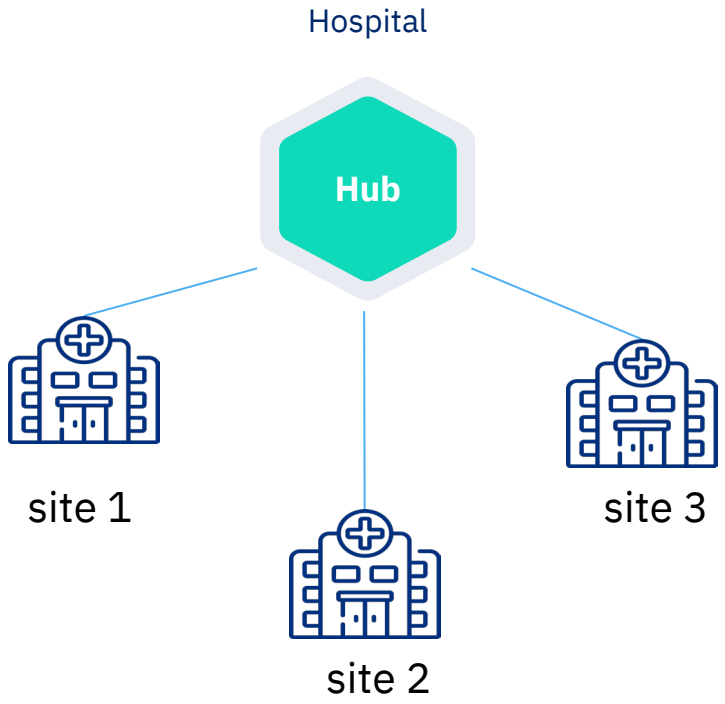

nutribullet

# Potential Solution: Have Every Data Source Use the Same Neutral (De-ID) Engine



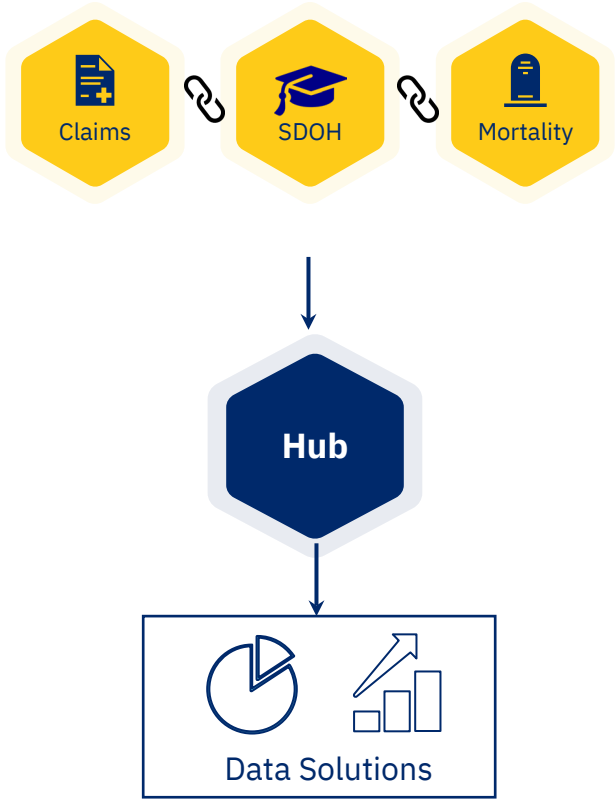
# Tokenization allows linking of a patient's records to build a longitudinal view of their journey

**Connect and Add Sites**



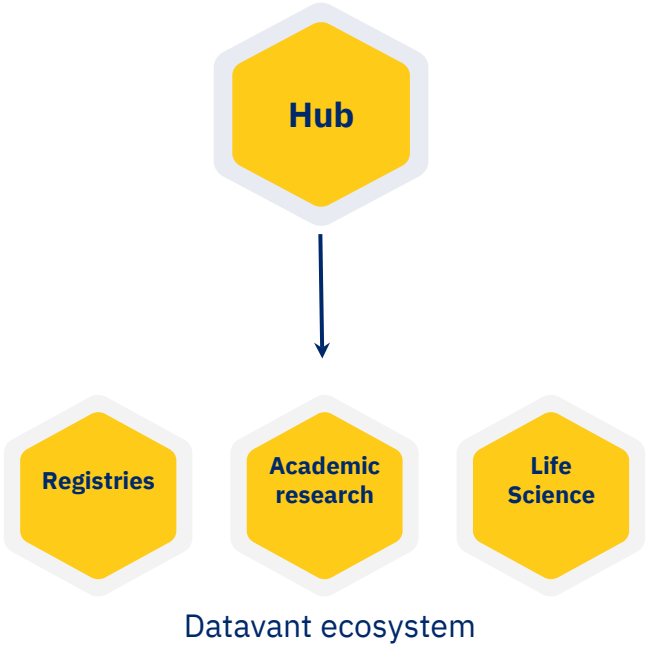

De-identification using **privacy-preserving linkages** unlocks possibilities for enrichment and partnership

**Data Enrichment & Discovery**



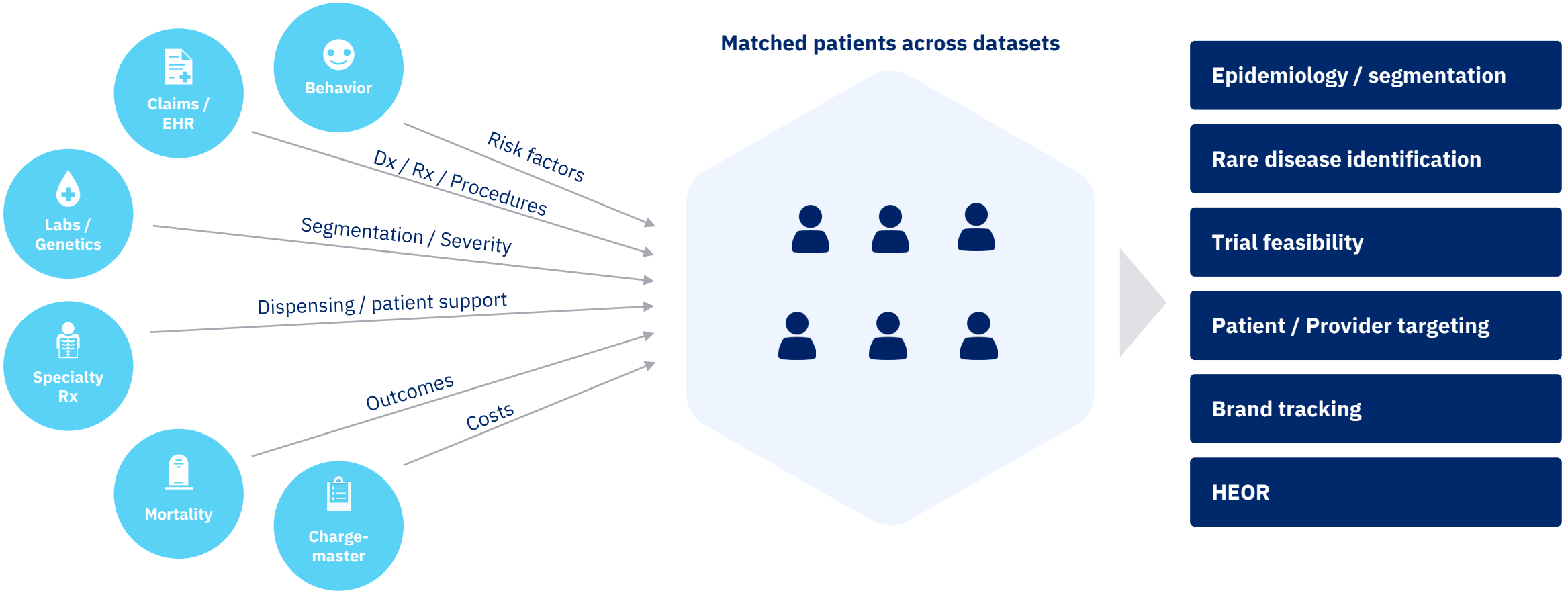

Discover relevant data partners or cohorts, **enrich data** with ecosystem partners

**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…



# 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

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


**80 million+ individuals**

**Longitudinal data 2009-2023**

**8 clinical networks, 2 health plans**

**66 health systems**

**337 hospitals**

**1,024 community clinics**

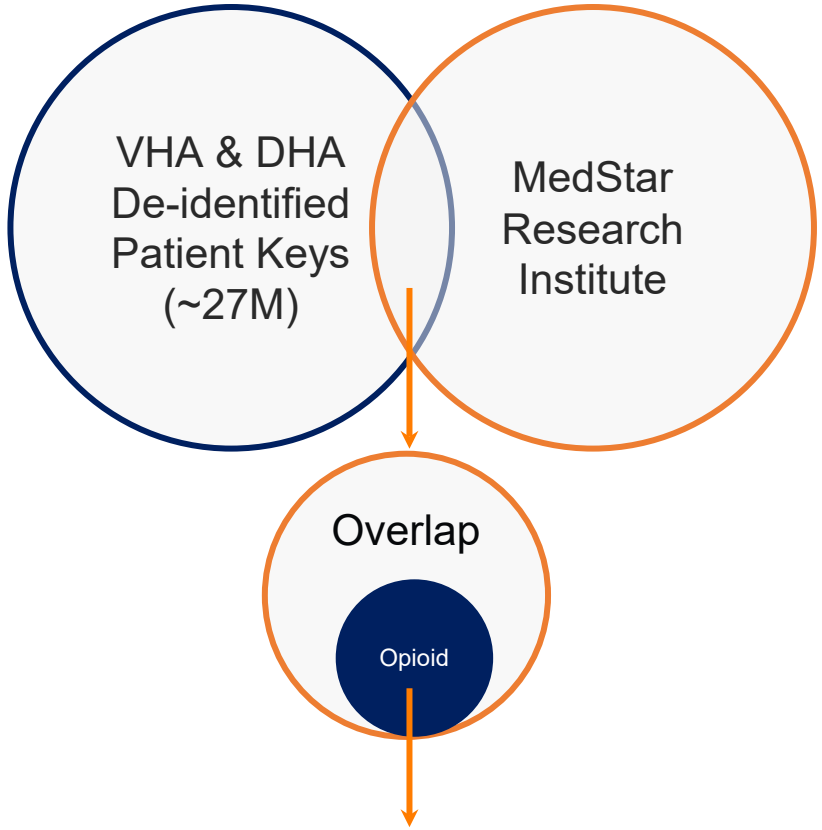
**3,564 primary care practices**

**338 emergency departments**

# DaVINCI Linkage with MedStar – Analysis Underway

De-identified linkage to ***discover*** shared patient cohort without revealing patient identity.
Method to formulate longitudinal patient record by seeking relevant data external to VA.

DaVINCI consists of a joint VA & DoD population

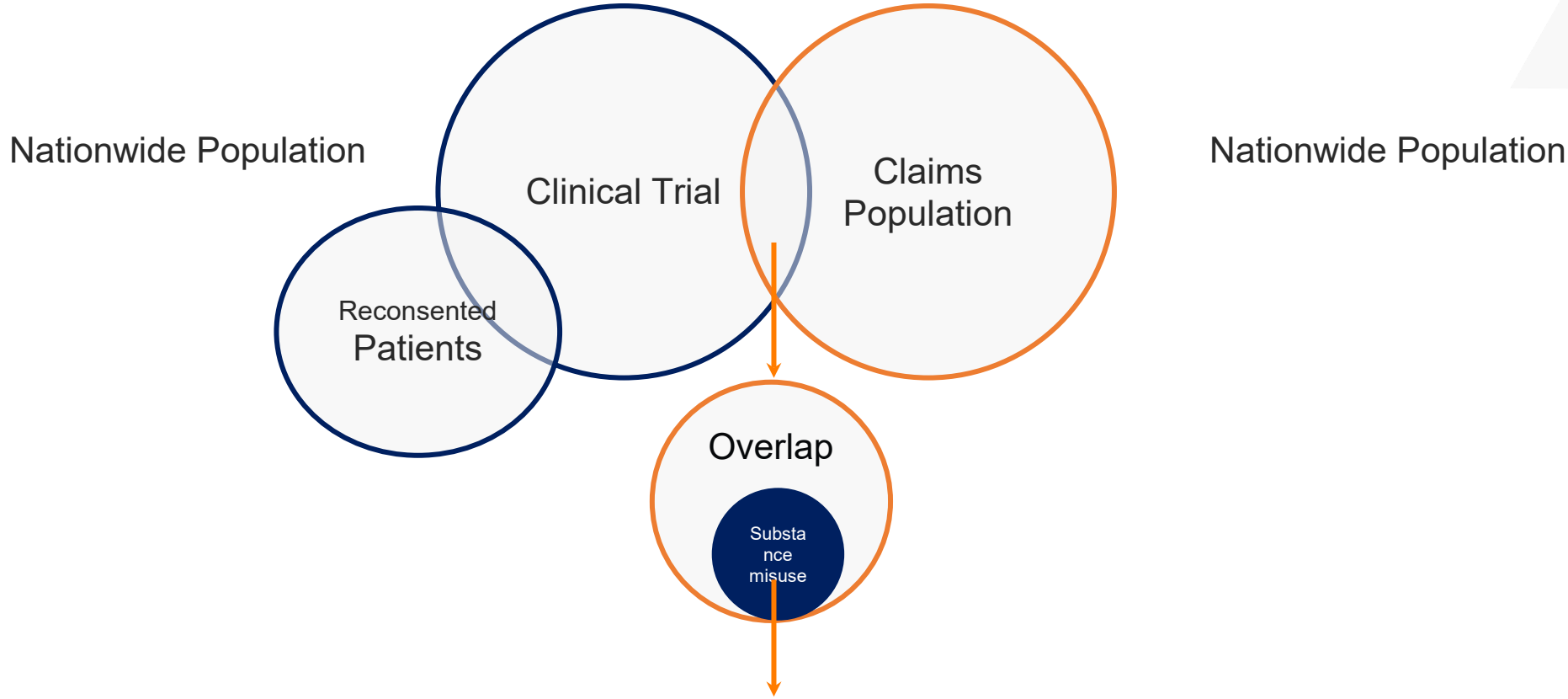



Populations from Washington, DC

**VA and Non-VA Hospital Patient Cohort**
**EHR data sent to VA**

# FDA Study using RWD on RWE – Analysis Underway

De-identified linkage to ***discover*** shared patient cohort without revealing patient identity.
Method to formulate longitudinal patient record by seeking relevant data.




**Academic Medical Center receives linked output**

COVID-19
RESEARCH DATABASE

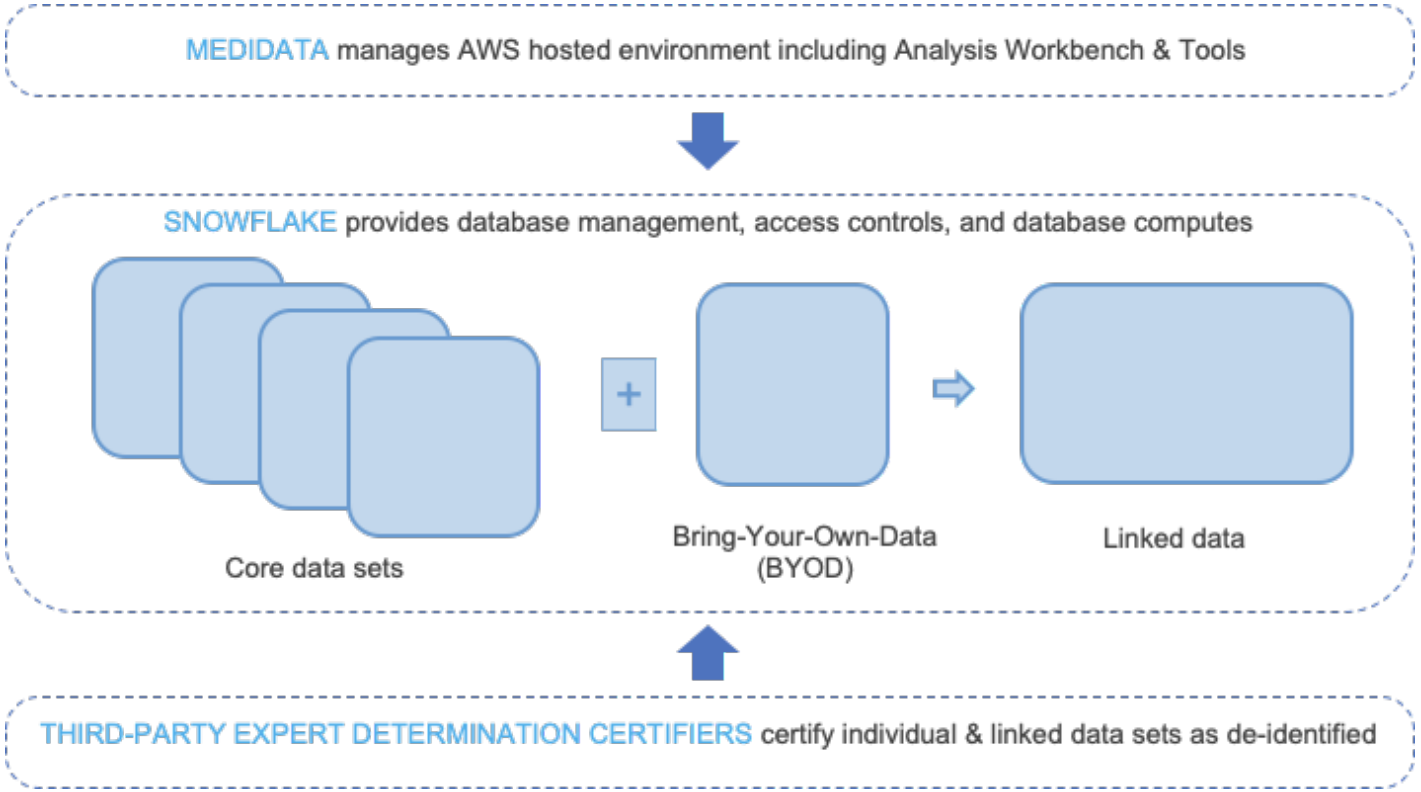
# Real World Linked Data Infrastructure established in response to the pandemic

>150 studies inflight  >400 researchers provisioned data

REAGAN-UDALL FOUNDATION for the Food and Drug Administration

INNOVATIONS IN Regulatory Science AWARDS

National claims
National mortality
Consumer demographic
Electronic health record
Nursing homes, SNFs
Diabetes monitoring
Life insurance
*and others*



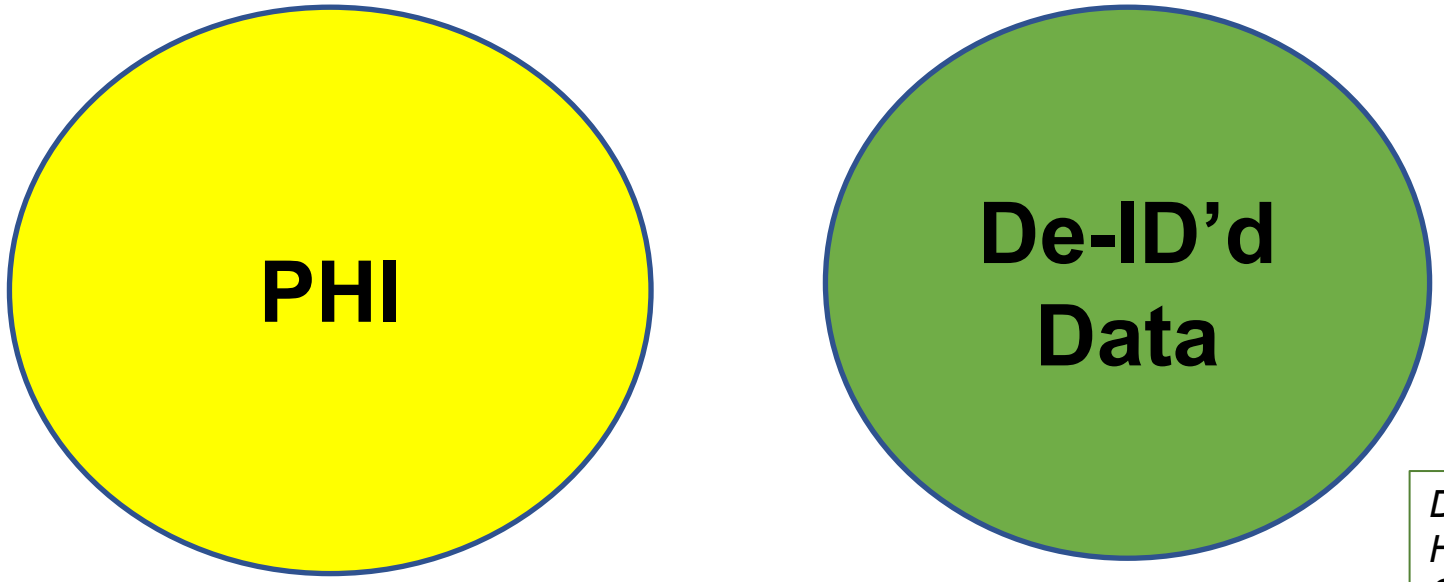

Certified de-identified
Preserve temporal info
Min necessary data
IRB exempt + waiver

Privacy+ Security Forum

# DE-IDENTIFICATION AND THE LAW(S)

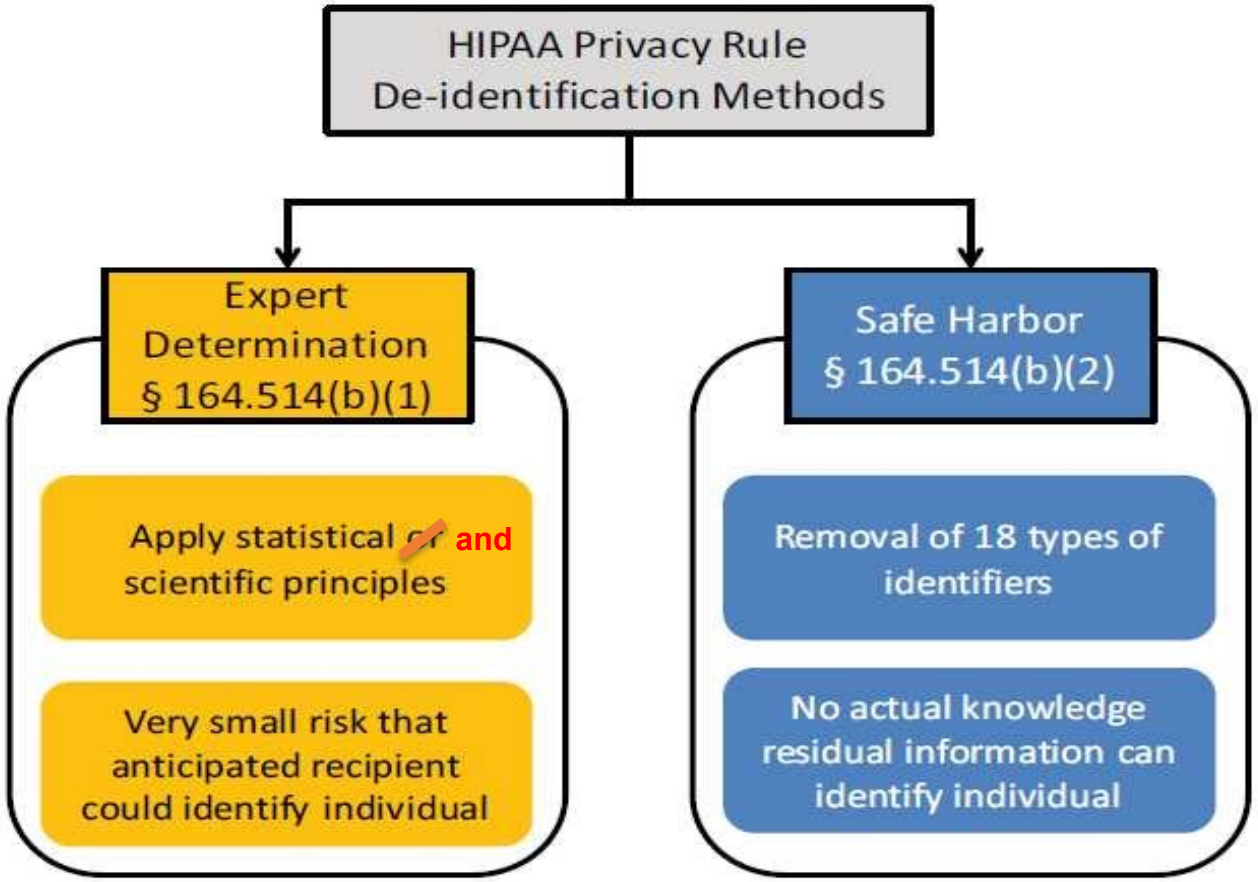

# De-identification under HIPAA - Basics

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*

# Two Methods of HIPAA De-identification

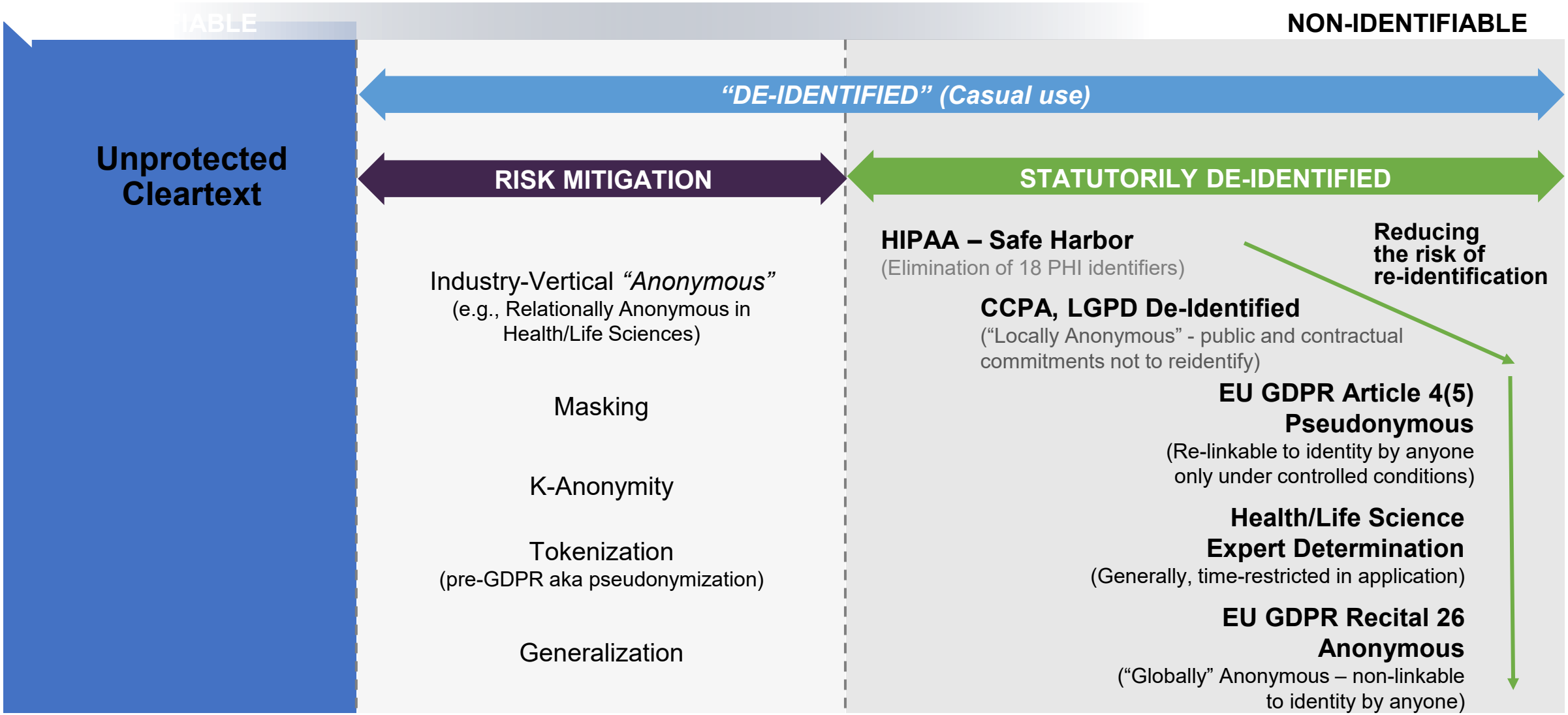



*ONLY TWO methods!*

Source: Office for Civil Rights (OCR) De-Identification Guidance (November 2012)

Corrected to match wording of §164.514(b)(1)

# Spectrum of Identifiable/Protected/Statutorily Non-Identifiable Data



- De-ID'd health data brings vast benefits to humanity – clinical trials, real-world evidence of treatment effectiveness, new tests and treatments, greater efficiency, scientific advances
- Achieving that HIPAA standard of de-ID'n is thus crucial to ecosystem of data liquidity. Massive investment by countless stakeholders to achieve and maintain HIPAA de-ID'n status. Standardization is key.
- *But….then along come new privacy laws. When they include novel and divergent de-ID'n definitions, that spells Complexity. Trouble.*

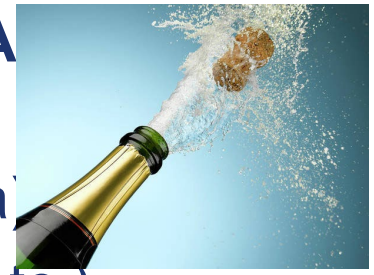
## CA CCPA (Original)

- 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 ID'n**
  - "Patient information" is broadly defined (PHI plus other medical data)
  - But does not include consumer digital health data (smart watches, etc.)



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


- ***Four new provisions apply to de-ID'd patient information**

## *Okay, that's CA. What about the other new state consumer privacy laws??*

All enacted to date (i.e., CA, CO, CT, IN, IL, VA, UT, and WA*)
(plus TN and WA, awaiting gubernatorial signature)
have a similar two-tier structure:

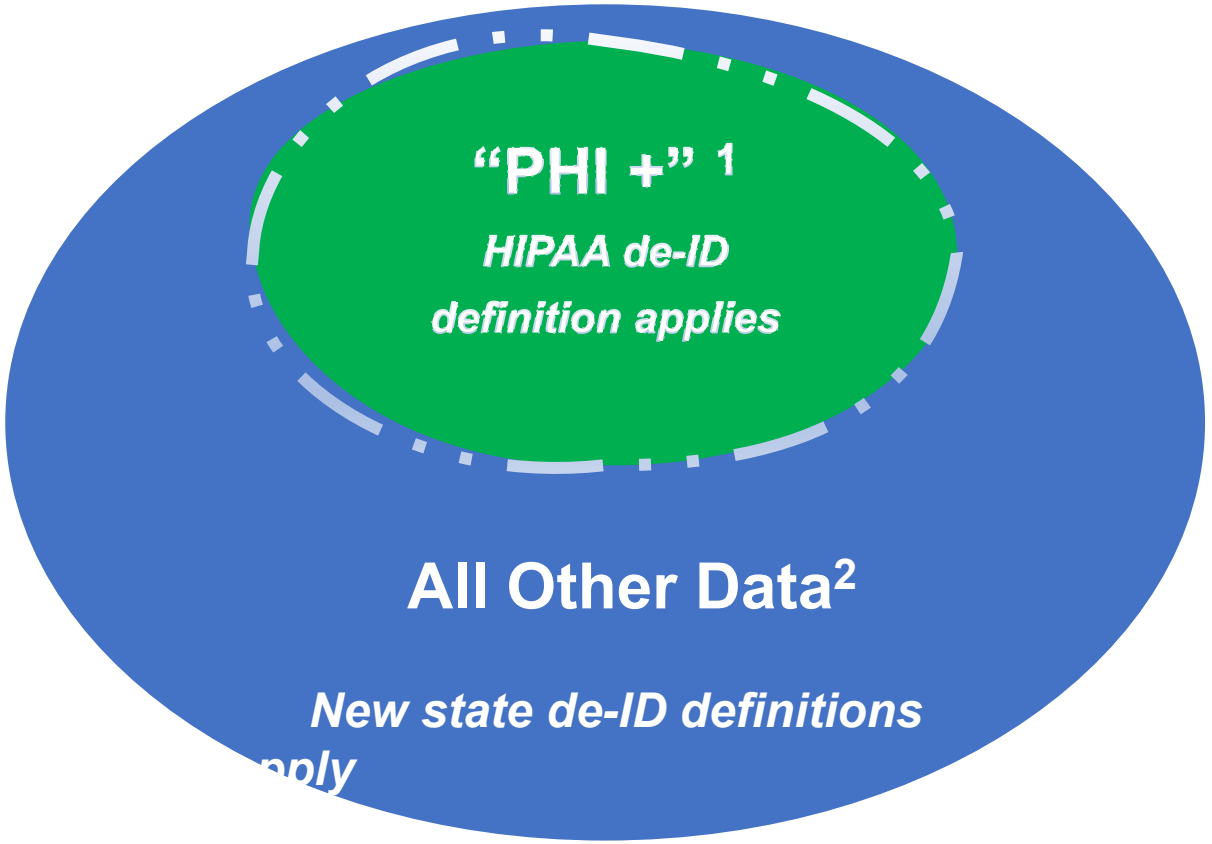
- **HIPAA de-ID'n applies to "PHI Plus" (PHI plus other medical data)**
- **New state-specific de-ID definition applies to all other data**

*Treating WA's new "consumer health" law as a general privacy law here due to its broad and unclear scope

# Which state De-ID standard applies to which data?

[1]**"PHI Plus"** is "patient information" in CA law and has other designations under other state laws. In essence, it 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.



## ***More complexities with de-ID'n under the new state laws***

- 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

Privacy+ Security Forum

# 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).

## *Audience Question*

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

## *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*

*How do you think that compliance with all the varying U.S. de-ID'n standards can be achieved?*

*I.e., how can a health data company substantiate that it has met all applicable federal and state de-ID'n standards?*

## New State Requirements 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**

Privacy+ Security Forum

## Audience Poll

***Time for a federal ban on re-identifying de-identified health data?***

Privacy+ Security Forum

# New State Requirements 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

# New State Requirements 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.

## 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

## 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
- *Again – the problem is inconsistency – not all new state laws recognize pseudonymization*

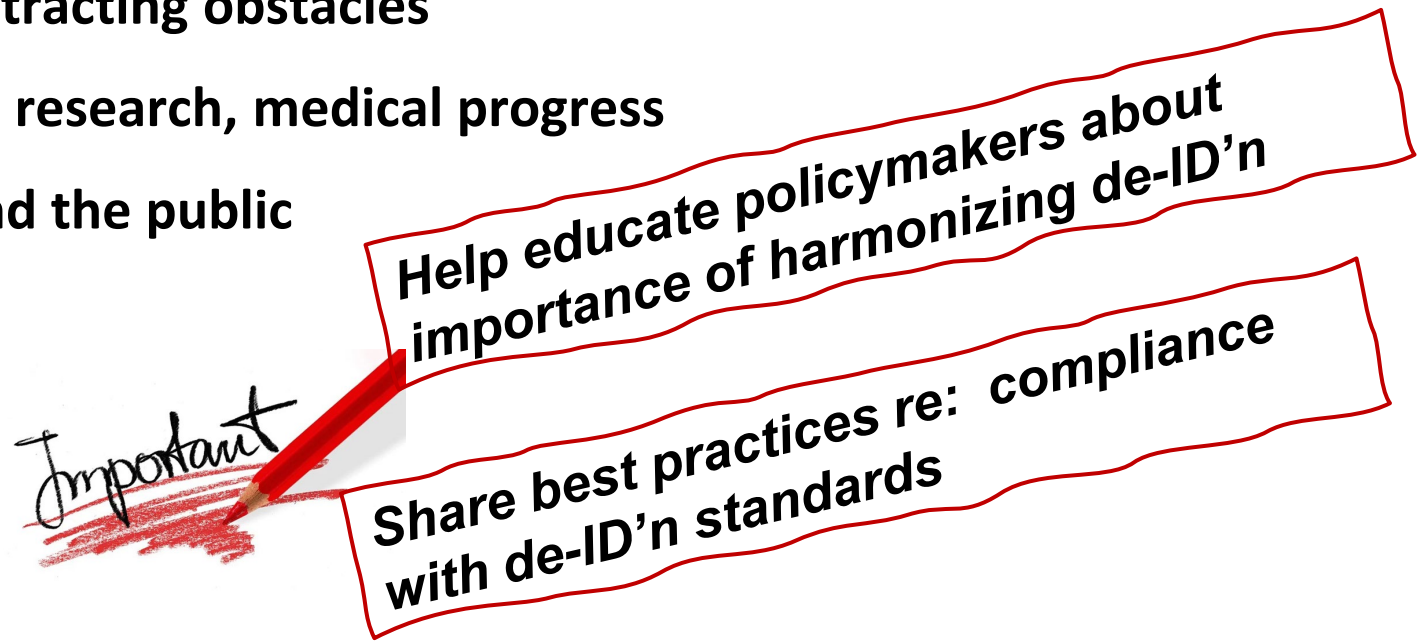
## New Oversight Duties re: De-ID'd Data

- Numerous new laws include a brand-new duty on data controllers to exercise oversight of entities to which they've disclosed de-ID'd data to monitor the recipients' compliance with their contractual commitments
- Compliance burden
- And –*surprise! – the wording of these oversight duties varies by state. Some apply to users of de-ID'd data; some to recipients*

## 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

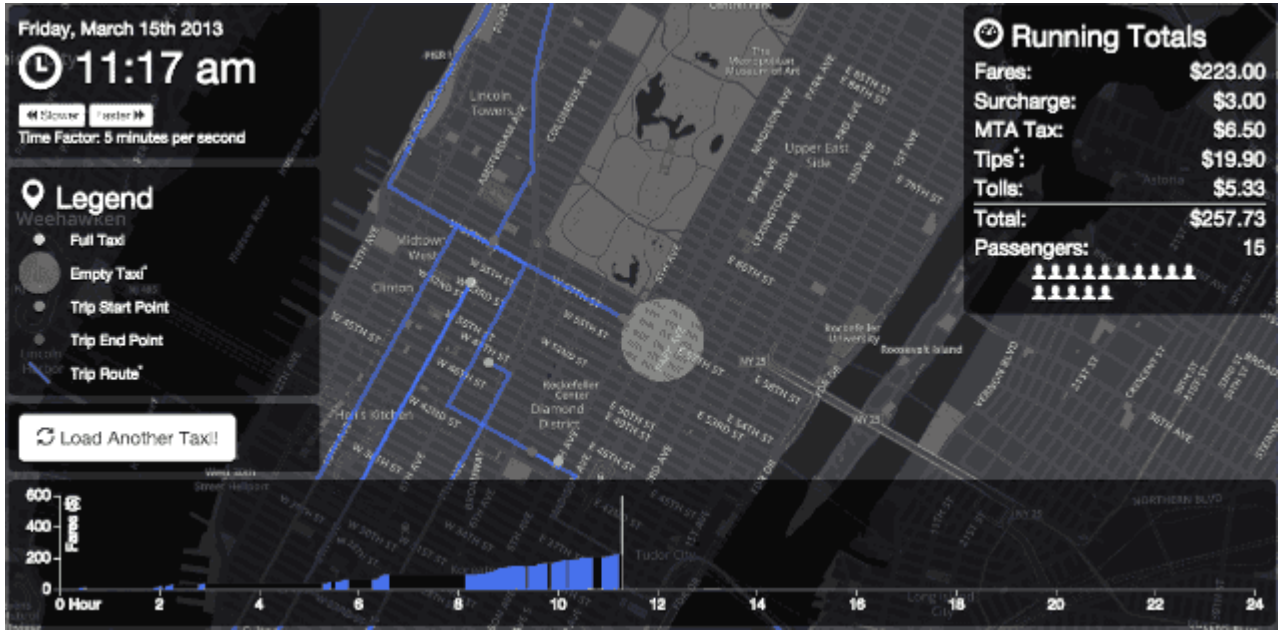
# Data Practitioner Perspectives

## Why De-identified Data?

- HIPAA – “Privacy Rule does not restrict the use or disclosure of de-identified health information, as it is no longer considered protected health information.”[1]
- GDPR – “The principles of data protection should therefore not apply to anonymous information…”
- CA (CPA) – "Personal information does not include information that is publicly available or that is de-identified.”
- Brazil (LGPD) – “Anonymized data shall not be considered personal data…”

[1] *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. HHS (Nov. 12, 2012

# Fear and Loathing: NY Taxi and Limousine Commission




"…city officials had attempted to anonymize certain identifying details associated with every ride - namely the medallion number, an alphanumeric code assigned to each taxi cab in operation, and the hack license number, which is assigned to drivers authorized to operate a yellow taxi…by running both sets of numbers through a notoriously weak cryptographic algorithm known as MD5."

**J.K. Trottor, Gawker, Public NYC Taxicab Database Lets You See How Celebrities Tip (2014)**

Privacy+ Security Forum

## “de-identified” vs “De-identified to a standard”

- "used a cipher”
- “removed name”
- “shifted dates”
- “can we get an exception?”
- “created our own process”

Privacy+ Security Forum

## Questions

- What are pros and cons of Safe Harbor vs. Expert Determination methods of HIPAA de-identification?
- What to consider when deciding between a Limited Data Set (LDS) vs De-identified data?
- What are the unique concerns or issues to consider when integrating consumer or commercial data with De-identified data?

Privacy+ Security Forum

## Questions

- What to consider when evaluating data retention issues with De-identified data?
- What are ways to consider employing De-identified data to reduce privacy risk?
- How do you know if you can de-identify data?

Privacy+ Security Forum

## Questions

- What about synthetic data?
- What are some of the residual risks of sharing de-identified data?
- What are hidden benefits of de-identification?

# Winning cool awards

— APRIL 27, 2023

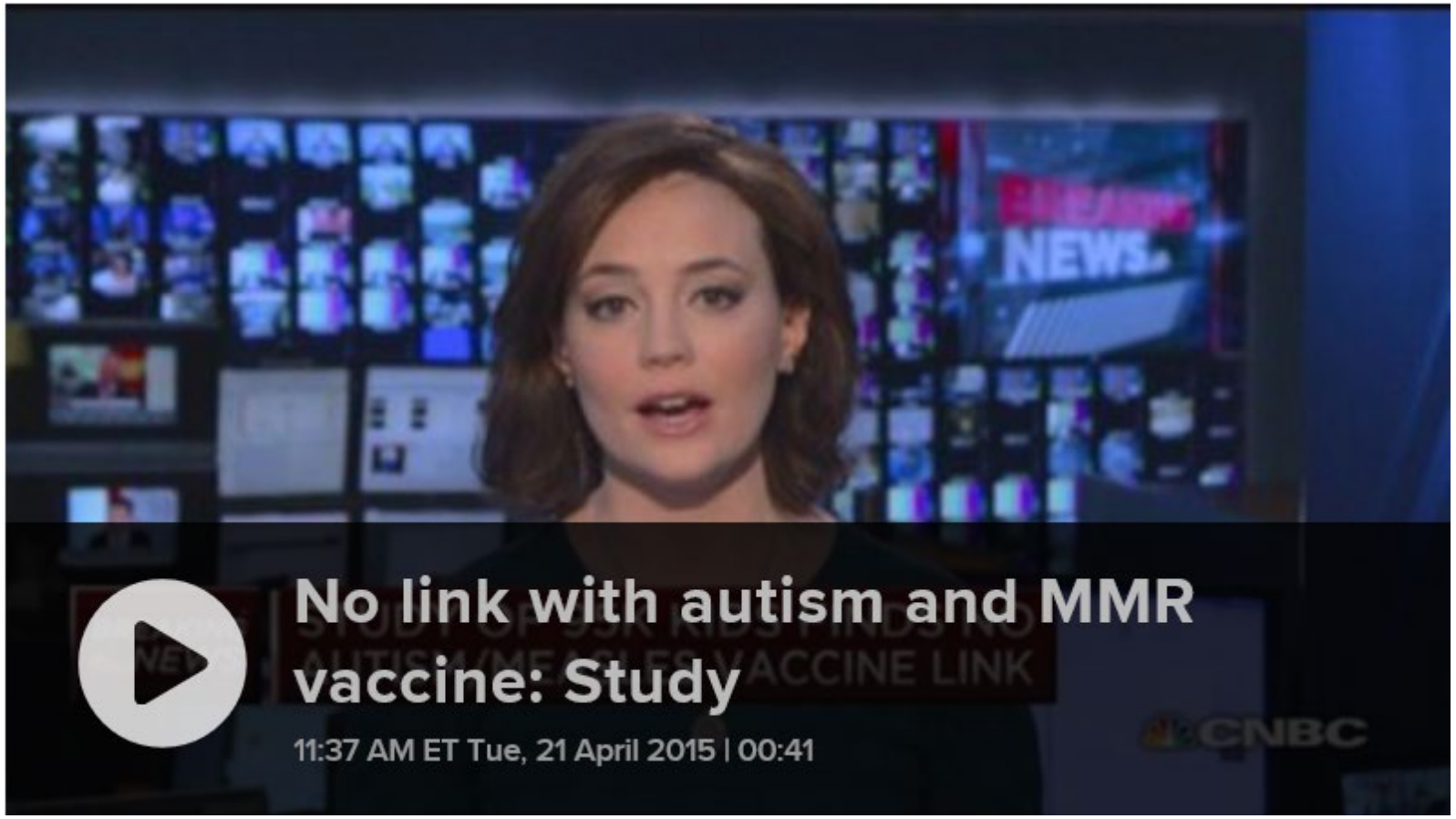
Today, the Future of Privacy Forum (FPF) — a global non-profit focused on data protection headquartered in Washington, D.C. — announced the winners of the third annual Award for Research Data Stewardship.

**NANCY LEVESQUE**
*Communications Director*
ABOUT NANCY →
BLOGS BY NANCY →

FPF is a long-standing advocate for privacy-protective data sharing by industry to the research community to advance scientific insights and drive progress in medicine, public health, education, social science, and many other fields. FPF established the Award for Research Data Stewardship in 2020 to recognize companies and academics that demonstrate innovative approaches and best practices for sharing private, corporate data to advance scientific knowledge.

With the third-annual Award for Research Data Stewardship, FPF honors two teams of researchers and corporate partners for their commitment to privacy and ethical uses of data in their efforts to help with emergencies related to diseases and natural disasters. The winning team is a collaboration between the Mayo Clinic researchers led by Rozalina McCoy, MD, MS, and health services company Optum. The honorable mention is a collaboration between Assistant Professor Xilei Zhao, PhD, at the University of Florida and location intelligence company Gravy Analytics. These partnerships were awarded based on the strength of their research, adherence to privacy protection in the sharing process, and the company's commitment to supporting academic research.

# Making healthcare better



## Reference Material

- HOW DATA CAN BE USED AGAINST PEOPLE:
- A CLASSIFICATION OF PERSONAL DATA MISUSES https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3887097

- Optum Autism Study: https://jamanetwork.com/journals/jama/fullarticle/2275444

- FPF Award Study: https://grantome.com/grant/NIH/K23-DK114497-02#:~:text=Severe%20hypoglycemia%20and%20hyperglycemia%20are%20often%20preventable%2C%20yet,and%20reliable%20means%20to%20identify%20high%20risk%20patients

# Questions + Contact

Privacy+ Security Forum


## Daniel Barth-Jones, PhD

**Principal Privacy Expert**
**Privacy Hub by Datavant**

danielbarth-jones.privacyhub@datavant.com



## Ann Waldo, JD

**Waldo Law Offices**

awaldo@waldolawoffices.com



## Peter Dumont

**Chief Privacy Officer**
**Optum Labs**

peter.dumont@optum.com



## Claire Manneh, MPH

**Head of Provider Partnerships**
**Datavant**

claire@datavant.com

Privacy+ Security Forum

# *Reference Slides*

# 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—

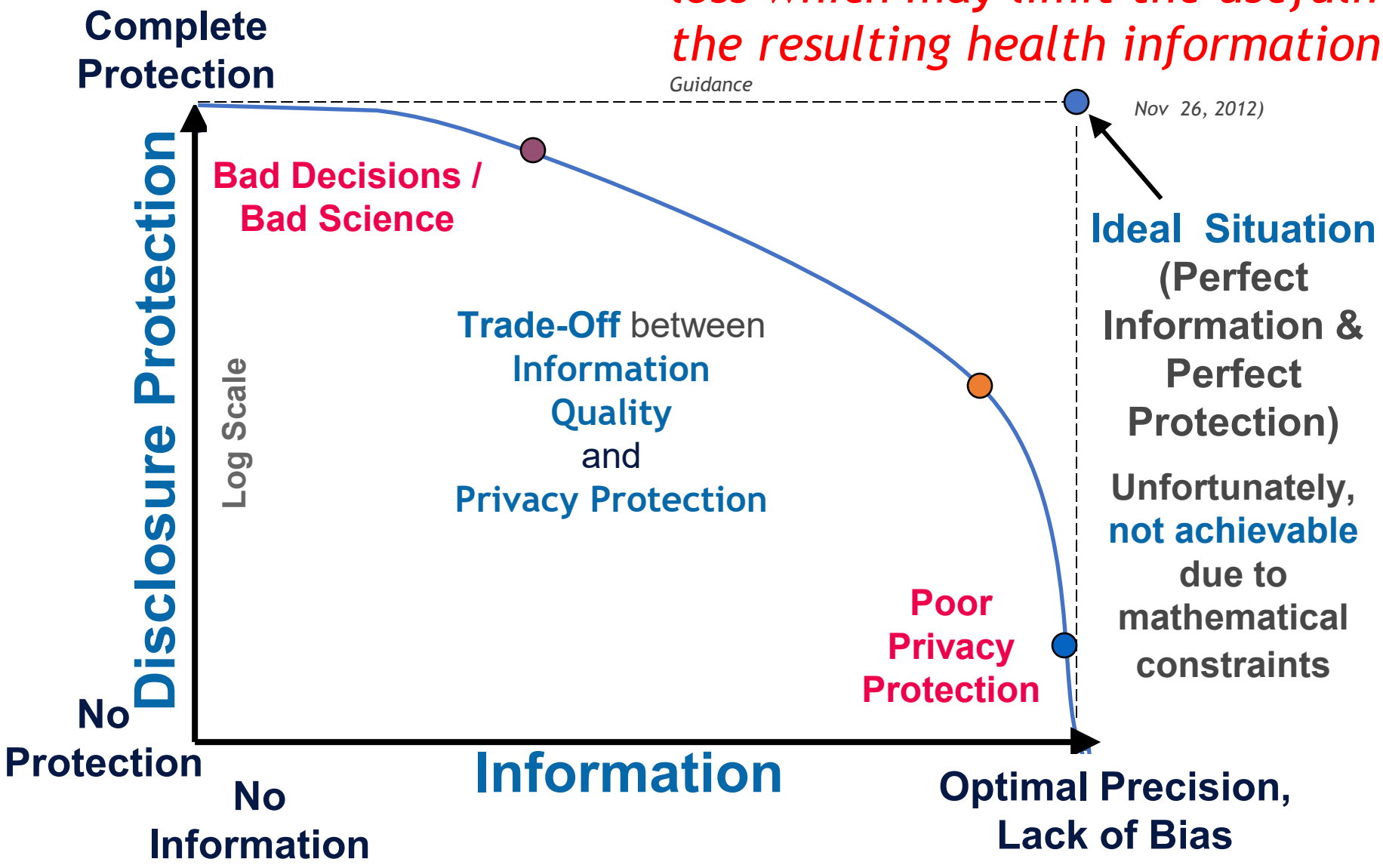
(i) to comply with all of the provisions of this paragraph with respect to the information; and

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

# More nuances re:  de-identification

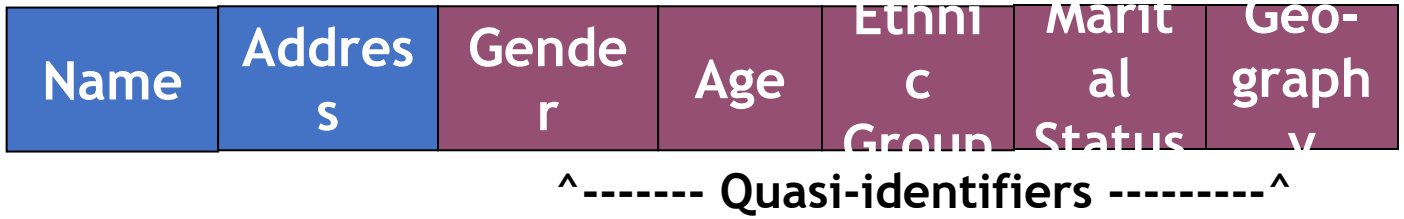
# The Inconvenient Truth:

*"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)*



## *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.).

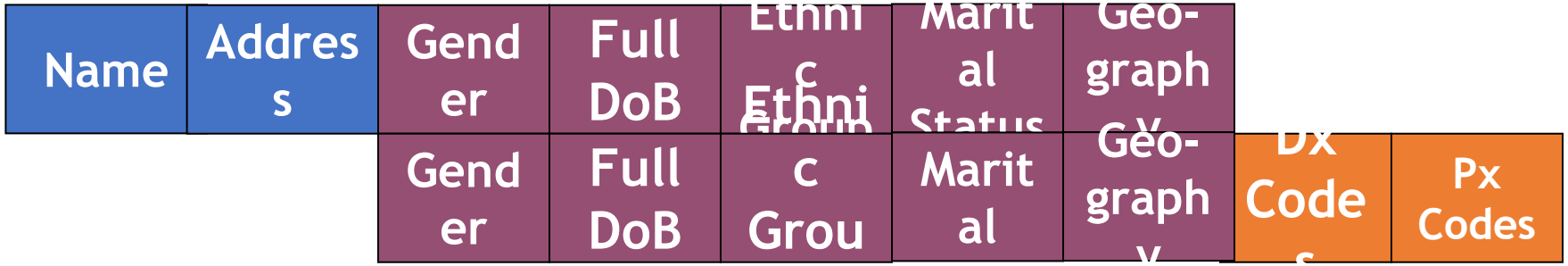
Note that this includes even fields that are not "PHI".

# *Key Resolution*

Key "*resolution*" exponentially increases with:

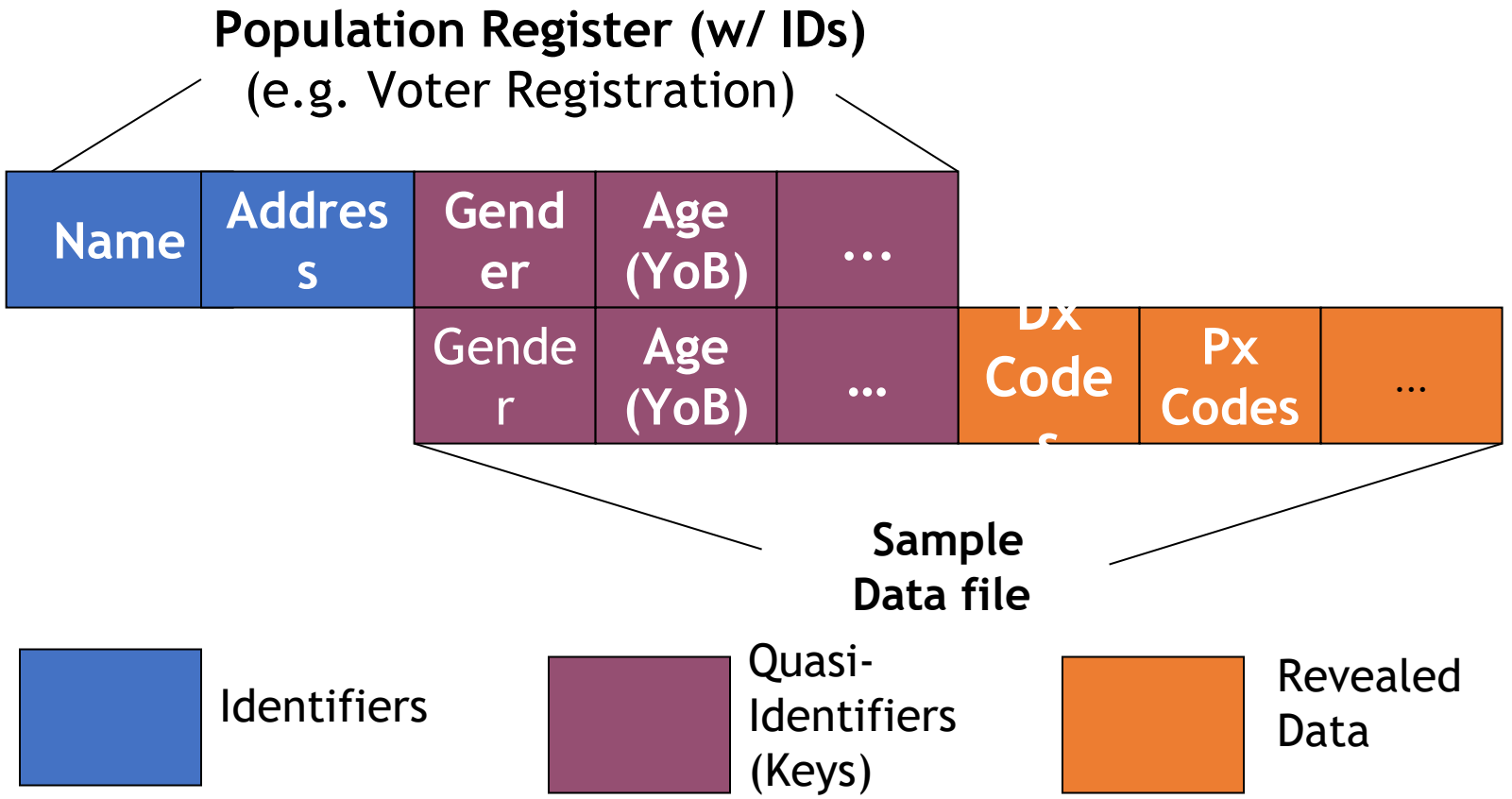
1) the number of matching fields available

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



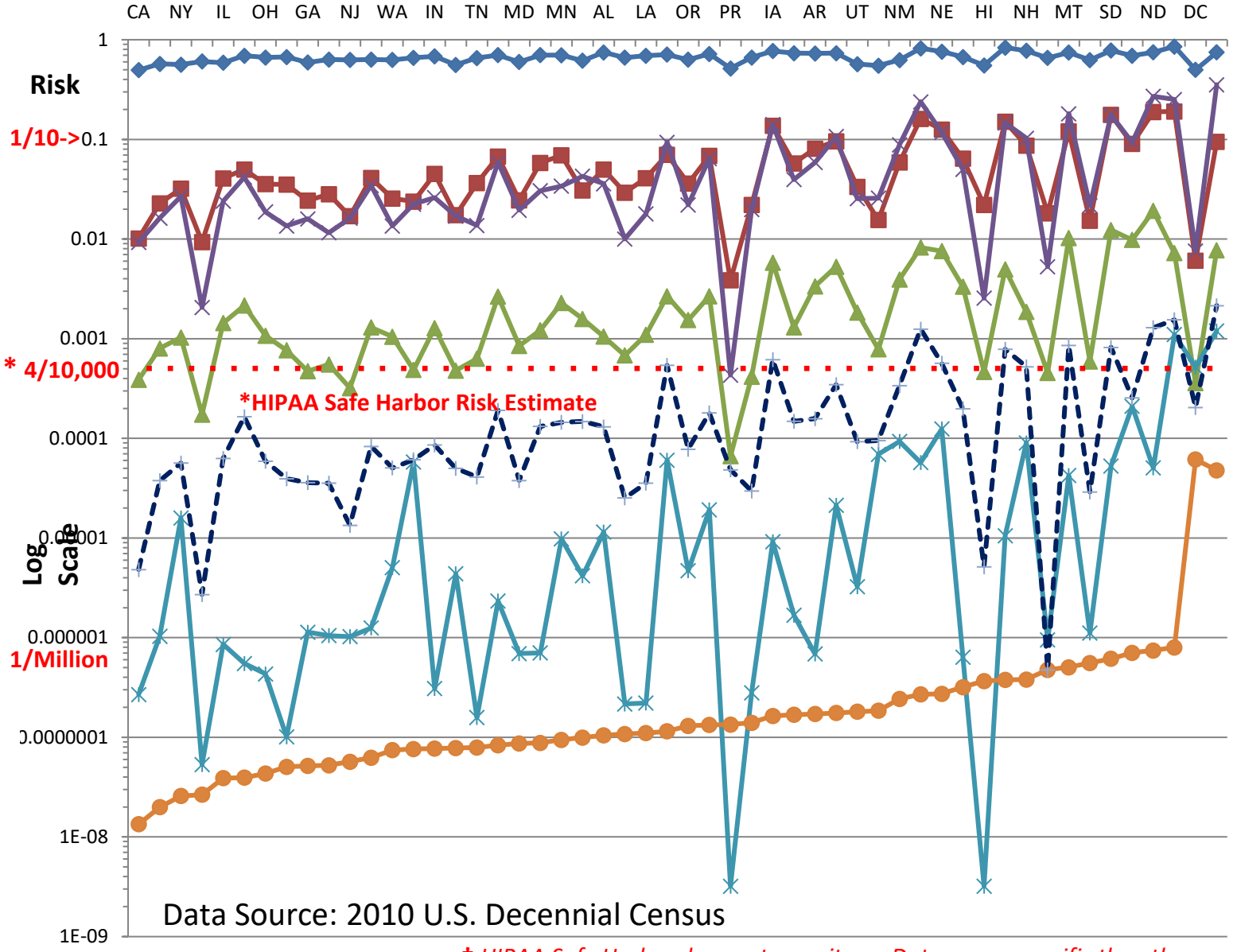
# *Record Linkage*

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



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

*(States ordered by Population Sizes)*




***Combined Quasi-Identifier Legend***

*DoB = Date of Birth*
*MoB = Birth Mnth & 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*

Not Safe Harbor Compliant

- DoB,Z5 †
- MoB,Z5
- YoB,Z5
- DoB,Z3
- MoB,Z3

Safe Harbor

- YoB,Z3
- YoB,Z3,Race

*Graph © DB-J 2013*

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

# *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 matrixes, 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

# 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.***

# 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)

# Recommended Skills for De-Identification Expert Teams

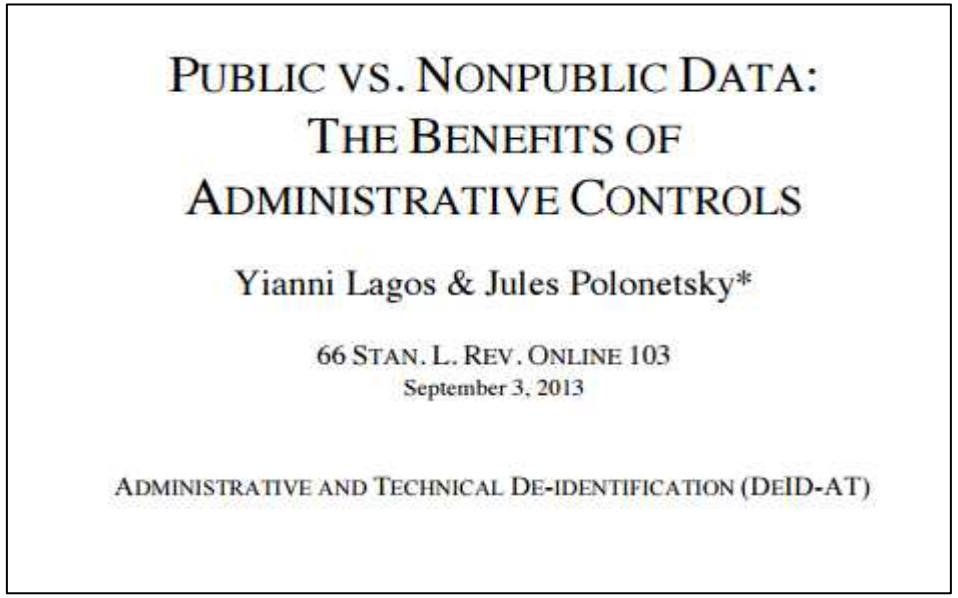
- Statistical Disclosure Limitation/Control Theory & Practices
- Privacy Preserving Data Publishing and Mining
- HIPAA/HITECH and Data Privacy Law
- Corporate Compliance and Data Governance
- Medical Informatics and Medical Coding/Billing Systems
- Biostatistics/Epidemiology
- Geographic Information 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

# 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?*

# *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)

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) – Should it be applied nationally?***

# 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 timeframe.

# 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).

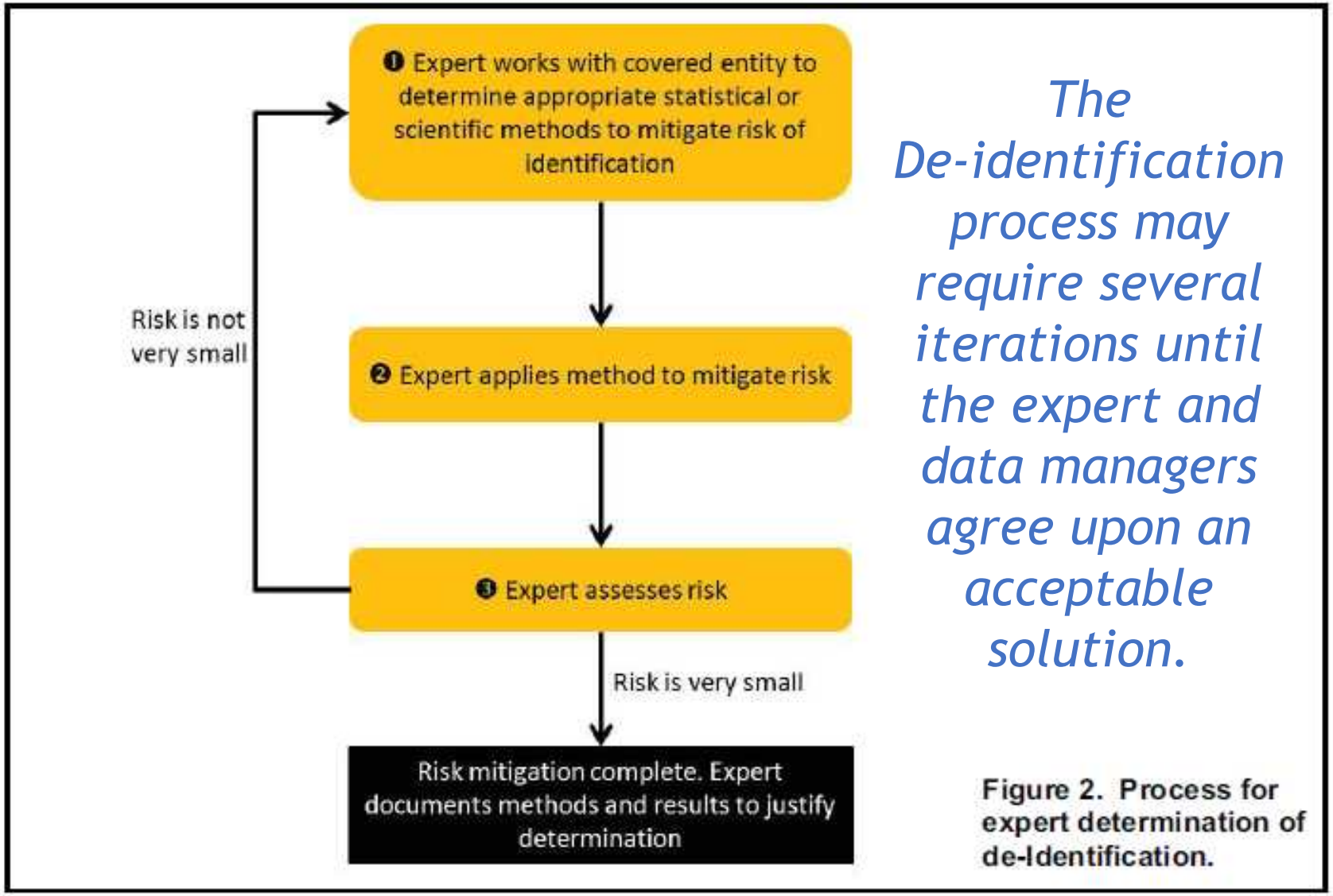
# Q2.5 *Can an expert derive multiple solutions from the same data set for a recipient?* *(Cont'd)*

- 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.

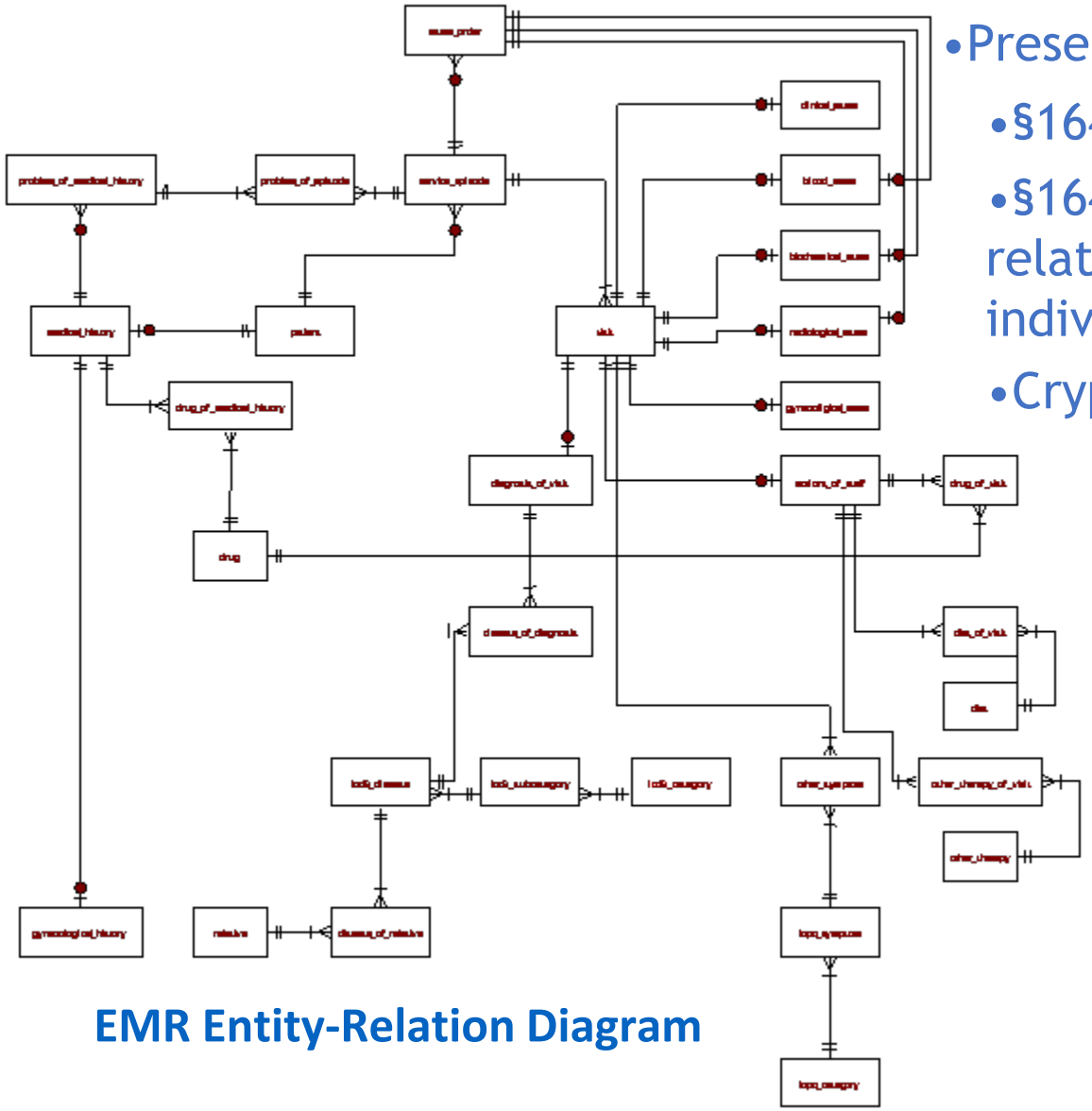
# Q2.8. *What are the approaches by which an expert mitigates the risk of identification?* *(Cont'd)*

- 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.

# 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