March 23, 2022

# De-identification Workshop

Ann Waldo, JD
Waldo Law Offices

Allison Bender, JD
Denton's

Daniel Barth-Jones, PhD
Assistant Professor of Clinical Epidemiology,
Mailman School of Public Health, Columbia University

Khaled El Eman, PhD
Professor, University of Ottawa
SVP and General Manager, Replica Analytics

Privacy+ Security Forum

Privacy+ Security Forum

## *LAW*

- Health data de-identification under HIPAA and new laws – harmonized or divergent standards?
- New state law requirements re:  de-identified health data
- New de-identification definitions for non-health data
- Real-world challenges for de-identified data in commercial data transactions and M&A deals
- Regulatory, litigation, and legislative risks

## *TECHNOLOGY AND STATISTICS*

- Measuring disclosure risks
- Multi-party data set linking
- Differential privacy
- Synthetic data

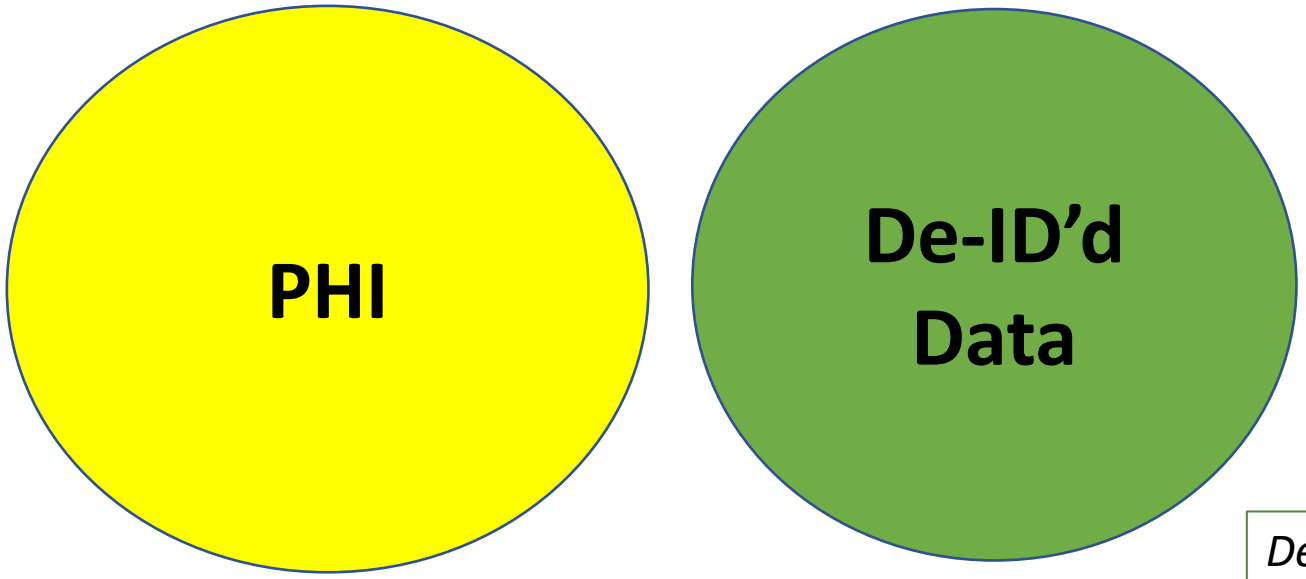
**Ann Waldo**
Principal,
Waldo Law Offices, PLLC

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.

Privacy+ Security Forum

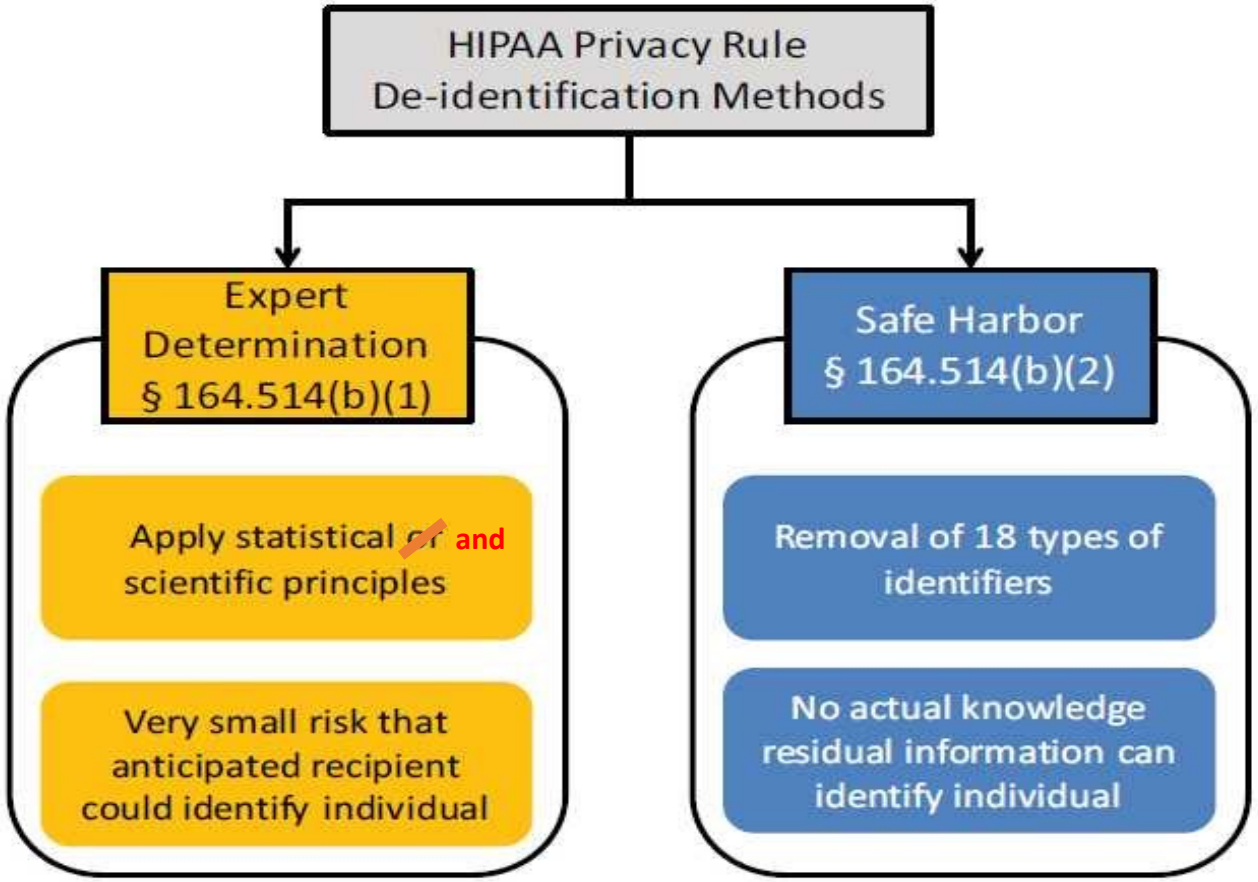
## De-identification under HIPAA - Basics

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




*De-ID'd data is outside HIPAA*
*Contract restrictions may apply*

# 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)

- 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. If they include novel and divergent de-ID'n definitions, that spells Trouble.*

## CA CCPA (Original)

- Brand new definition of “deidentification,” bearing no resemblance to 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 for health data under HIPAA and CCPA
    - *Successful!*
    - *Multi-stakeholder collaboration, including privacy advocates*
    - *CA AB 713 (2020)*

## De-ID'n under CA Law Today*

- **De-ID'n for patient information in CA now harmonized with HIPAA de-ID'n**
  - "Patient information" is broadly defined
  - But does not include consumer digital health data (smart watches, smart scales, etc.)
- **Four new provisions apply to de-ID'd health data**


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


**Under both present law (CCPA, including AB 713) and CPRA (eff 1/1/2023)*

Privacy+ Security Forum

## New CA Provisions Regarding De-ID'n

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

- Cannot re-identify, or attempt to re-identify, de-ID'd data that is exempt from CCPA because of newly harmonized de-ID'd definition
- Scope - a business or other person
- 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*

Privacy+ Security Forum

# New CA Provisions Regarding De-ID'n

## 2) Contractual Requirements for Sales

- Scope - one of the parties resides or does business in CA
- 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

# New CA Provisions Regarding De-ID'n

## 3) 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 CA Provisions Regarding De-ID'n

## 4) 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 health data, *but then became re-identified,* becomes subject to applicable privacy law, including HIPAA, CA CMIA, and CCPA, if applicable
- *My own view on the legal impact of this one*

## CCPA General Definition of De-Identification (Applies to All Data Except Patient Information)

"Deidentified" means information that **cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer**, provided that a business that uses deidentified information:

(1) Has implemented technical safeguards that prohibit reidentification of the consumer to whom the information may pertain.
(2) Has implemented business processes that specifically prohibit reidentification of the information.
(3) Has implemented business processes to prevent inadvertent release of deidentified information.
(4) Makes no attempt to reidentify the information.

*Qs – What if a business de-ID's data for one purpose but then re-ID's for a legitimate purpose? Does this mean it was never de-ID'd in the first place? What about de-ID'd information responsibly released to the public?*

**Note that CPRA alters the above definition by including references to inferences**

## Other New State Privacy Laws

***The outstanding news for medical research and health data* – so far, all the new comprehensive privacy laws have harmonized their de-ID'n definitions for health information**

- VA Consumer Data Privacy Act
- CO Privacy Act
- Utah Consumer Privacy Act

***But for general (non-health) de-ID'd data*** - the definitions in CA, VA, CO, and UT are already diverging

***And definitions diverge further in pending state and federal bills***

***Divergent definitions of de-ID'n pose real challenges to data interoperability and fluidity***

# Speaker

**Allison Bender**
Partner, Dentons US LLP
Washington, DC
D +1 202 496 7362 | M +1 703 853 7999
allison.bender@dentons.com

Allison J. Bender is a partner in Dentons' Venture Technology and Emerging Growth Companies practice. Allison's practice focuses on helping clients strategically manage privacy and cybersecurity risk, including as it relates to digital health.

She advises companies in developing and implementing practical information governance programs that "fit" the organization, taking into account the nature of their data, systems, industry, and stage, as well as the laws of the jurisdictions in which they do business. She guides efforts to maximize the impact of privacy and security policies, procedures, and assessments. She has developed and facilitated executive-level incident response exercises, conducted enterprise-wide risk assessments, aided companies in preparing for certifications and audits, overseen legal issues in targeted penetration testing, and provided counsel on responses to reported vulnerabilities. Clients turn to her for advice on cutting edge legal issues in high-risk vendor contracts and in developing new products and services that raise privacy-by-design and security-by-design issues. She also has deep experience advising digital health, medical device, pharmaceutical, software, and other technology companies at all stages of maturity.

Allison is a noted public speaker in the field, including speaking at Mandiant Cyber Defense Summit 2019 and 2021, Black Hat USA 2018, and the International Association of Privacy Professionals 2018 Global Summit. She also serves an adjunct professor at Georgetown University Law Center, teaching courses on cybersecurity law and national security regulation. Allison's prior experience includes government service as a senior attorney at the US Department of Homeland Security.

# Key Sources of Requirements for Data Use

- **Statutes and regulations**
  - International
  - US
    - Federal
    - State
- **Contracts, including NDAs and BAAs**
  - Customers
  - Vendors
  - Partners
  - Insurance
  - Investors
- **Individual authorization**
- **Informed consent**
  - Primary research purpose
  - Secondary research purpose
  - Other
- **“Reasonable” practice**
  - Industry standards and frameworks
  - Government guidance and policy statements

# Key Sources of Requirements for Data Use

- **Statutes and regulations**
  - International
  - US
    - Federal
    - State
- **Contracts, including NDAs and BAAs**
  - Customers
  - Vendors
  - Partners
  - Insurance
  - Investors
- **Individual authorization**
- **Informed consent**
  - Primary research purpose
  - Secondary research purpose
  - Other
- **“Reasonable” practice**
  - Industry standards and frameworks
  - Government guidance and policy statements

- **General consumer data**
  - Data breach notification laws
  - Information security laws
  - Records retention and disposal laws
  - Privacy rights legislation (e.g., CCPA, CPRA, others)
  - Consumer protection laws
- **Health information privacy**
  - Federal:
    - HIPAA (as amended)
    - SAMHSA
    - The Common Rule
    - GINA
    - FTC Act
    - Others
  - *State:*
    - HIPAA equivalents like CMIA
    - Insurance laws and regulations
    - Licensing requirements
- **Intellectual property rights**
- **Confidential information restrictions**

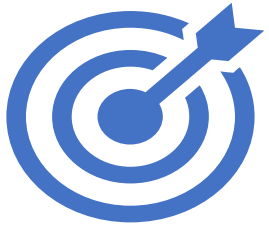
# Health Data Keywords

Privacy+ Security Forum

| | | | |
|---|---|---|---|
| Personal Information | Personally Identifiable Information | Personal Data | Protected Health Information |
| Research Data | Deidentified / De-identified | Anonymised | Pseudonymised |

## Use during the contract period

- **Primary purpose:**
  - Perform the Services
- **Common secondary purposes:**
  - Internal development purposes,
  - Data aggregation
  - De-identification
  - Marketing
  - Sale



## Use after the contract period

- Internal use
- Research purpose
- Assignment or transfer of corporate assets
- Limited commercial purpose
- Unlimited purposes, including sale of data

# Maximizing Data Use Rights in Contract Negotiations: HIPAA

- **Does HIPAA apply to the data or the services to be performed?**
  - Are you performing work or buying data from a health care provider, group health plan, or health care clearinghouse?
    - If not, is the entity you are working with a business associate of any of the covered entities above?
  - Will any of the services performed be on behalf of a covered entity or another business associate for a covered entity for a health care treatment, payment or treatment purpose?
  - Will any of the services be paid for or reimbursed by insurance or Medicaid/Medicare, or will eligibility for reimbursement be checked as a condition of receiving services?
  - Are exceptions for certain uses and disclosures available, such as activities preparatory to research?
  - Is individual authorization for use and disclosure something the customer has or would be willing to seek?

- **If HIPAA applies to you as a business associate or sub-business associate:**
  - Do you need the right to perform data aggregation services across different customers?
  - Do you need to (or want to) be able to de-identify the data for the services, for other purposes during the contract period, or other purposes after the contract period?
    - If you remove all of the data elements required to meet the HIPAA De-Identification Safe Harbor Method, is that “enough” for your data use strategy?
    - Can you commit to not seeking to re-identify the data?

# Maximizing Data Use Rights in Contract Negotiations: Research

- **If not subject to HIPAA, does the Common Rule apply?**
  - Does the data derive from an institution that receives federal funding of any kind?
  - If the research is fully funded by private funds and the institution does not receive any federal funding, did the research institution nevertheless apply the Common Rule voluntarily?
  - Does the data relate to military servicemembers, prisoners, pregnant women, fetuses, neonates, or children?
  - Does the data relate to any other "vulnerable" population that may be of heightened risk?
- **What does the informed consent and research protocol say about research data use?**
  - Is the scope of the informed consent sufficient for the primary research purpose?
  - Was the informed consent potentially invalidated by combination with any other language?
  - What secondary research uses are authorized or arguably within scope?
  - Are supplemental notices or consents required, and if so, is that a commercially reasonable undertaking?
- **Do you need access to, or retention of, biospecimens (e.g., blood, tissue, DNA) in addition to the research data?**
  - If depositing the specimen in a publicly available repository is required, is this consistent with your intellectual property rights strategy?
- **Do you need access to patient records or consumer digital health and wellness data in addition to the research data?**
  - If so, does the research protocol and informed consent permit (or at least not prohibit) combination of the research data with these other sources?

# Maximizing Data Use Rights in Contract Negotiations: CCPA

- **In the US, the California Consumer Privacy Act introduced new consumer rights for California residents, including the right to opt out of the "sale" of their personal information.**
  - "Sale" is defined very broadly.
  - The original CCPA language regarding deidentification was unclear and created uncertainties for covered entities and business associates.
- **AB 713 amended CCPA to clarify that PHI de-identified under HIPAA is considered "deidentified" under the CCPA. However, the following additional requirements apply:**
  - A business must disclose:
    - whether it sells or discloses personal information de-identified pursuant to HIPAA, and
    - the chosen HIPAA de-identification methodology (i.e., Safe Harbor or Expert).
  - Deidentified data later re-identified will again be subject to federal and state law, including HIPAA, the Common Rule, and CMIA.
    - HIPAA does not require covered entities to conduct ongoing monitoring
    - HIPAA would not apply to recipients of de-identified data, even if they re-identified the data; CCPA would
  - A business must include contract language that:
    - Bars reidentification, subject to certain purpose-oriented uses, with a flow-down requirement
    - Requires contracts for sale of de-identified health data to prohibit reidentification (for contracts on or after Jan. 1, 2021) if one party to the contract resides in or does business in the state (no non-profit exception, no thresholds)

# Potential Downstream Effects of Inadequate Consent/De-identification

- **"Hold yourself accountable—or be ready for the FTC to do it for you**," Elisa Jillson, an attorney in FTC's privacy and identity protection division (April 19, 2021 blog)

- **Businesses must "be careful with the data that powers their model**" or face potential deletion of algorithms created with tainted data:

  - FTC settled with **Everalbum**, regarding use of photos uploaded by app users to train its facial recognition album, misrepresenting the degree of control users had regarding the algorithm and their ability to delete photos and videos upon account deactivation. As part of the 2021 settlement, Everalbum was required to delete algorithms and models even *partially* created using information in violation of COPPA, even if the majority of the information used was legally collected.

  - The FTC's proposed settlement with **WW International, Inc.** (fka Weight Watchers) and its subsidiary similarly require destruction of all personal information in violation of consent requirements as well as any models or algorithms based on that information.

- If buying or selling de-identified data, these regulatory developments merit a stronger focus on the quality of such data, privacy and adequate consent, limitations of liability, and indemnification.

# Break

# Speaker

**Daniel Barth-Jones, MPH, PhD**
Assistant Professor of Clinical Epidemiology
Mailman School of Public Health, Columbia University

Dr. Daniel C. Barth-Jones is an Assistant Professor of Clinical Epidemiology at the Mailman School of Public Health at Columbia University. Dr. Barth-Jones received his Master of Public Health degree in General Epidemiology and Ph.D. in Epidemiologic Science from the University of Michigan. Dr. Barth-Jones conducts research and provides consultation regarding how to best protect the privacy and identities of entities within health information databases while simultaneously preserving the analytic accuracy of such healthcare data for statistical analyses. His experience conducting and managing statistical disclosure limitation operations and research has spanned more than 20 years, involving activities in both the healthcare information industry and in academia. He has conducted educational training and made scientific presentations on statistical disclosure limitation to persons representing state and national healthcare organizations, commercial healthcare and healthcare information companies, federal agencies, and academia. In March 2010, Dr. Barth-Jones was one of a select group of statistical disclosure experts invited by the HHS Office of Civil Rights to serve as a presenter and expert panelist for their Workshop on the HIPAA Privacy Rule's De-Identification Standard. He has also authored several peer- reviewed publications and a book chapter on statistical disclosure assessment and control. He has performed numerous HIPAA-compliant statistical de-identification analyses with associated HIPAA expert determinations.

# 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"

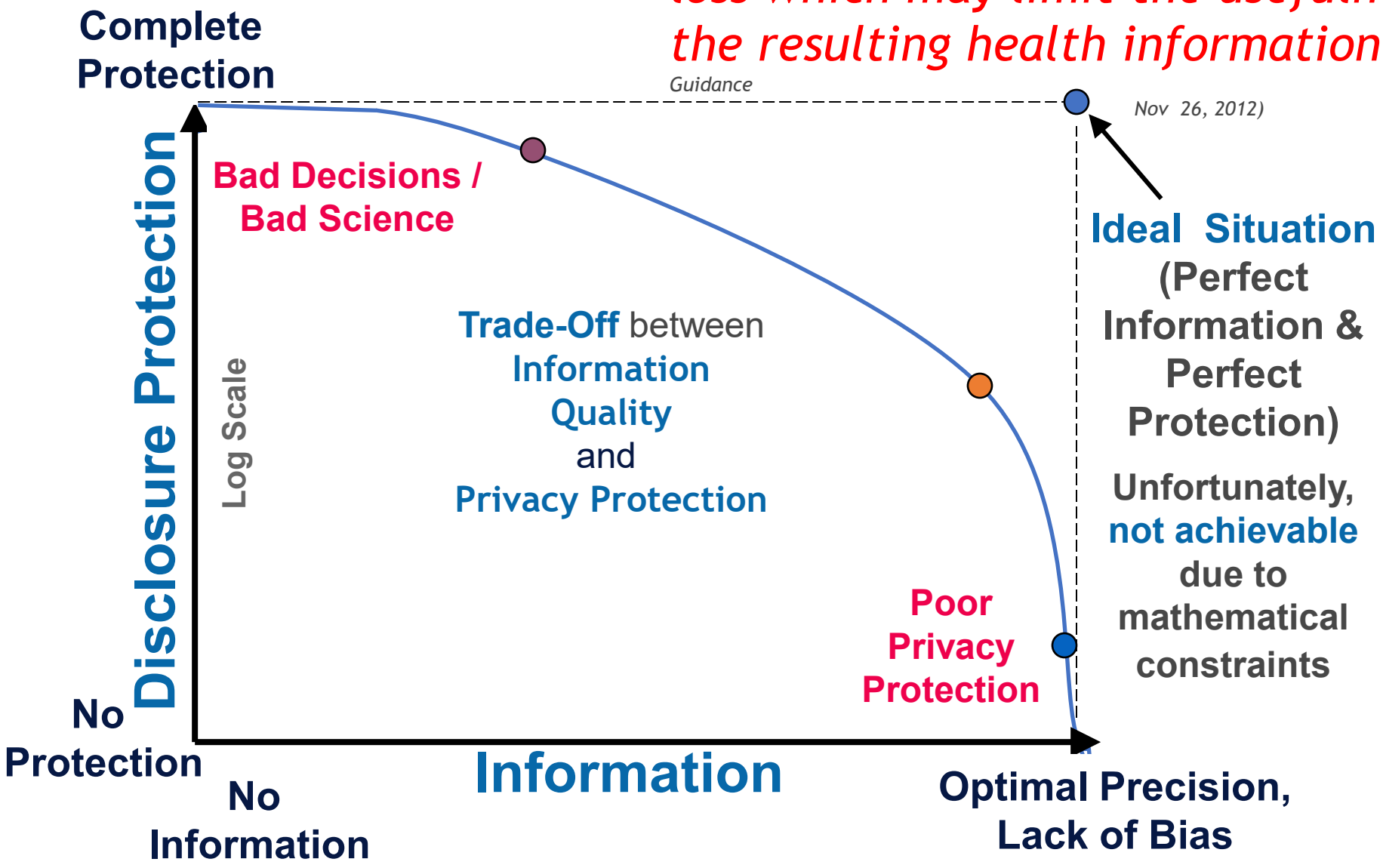
Health Information is not individually identifiable if:

*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)*



# Essential Re-identification Concepts

- Essential Re-identification and Statistical Disclosure Concepts
  - Record Linkage
  - Linkage Keys (Quasi-identifiers)
  - *Sample Uniques* and *Population Uniques*
- Straightforward Methods for Controlling Re-identification Risk
  - Decreasing Uniques:
    - by Reducing Key Resolutions
    - by Increasing Reporting Population Sizes

## *Quasi-identifiers*

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

| Name | Address | Gender | Age | Ethnic Group | Marital Status | Geo-graphy |
|---|---|---|---|---|---|---|

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

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

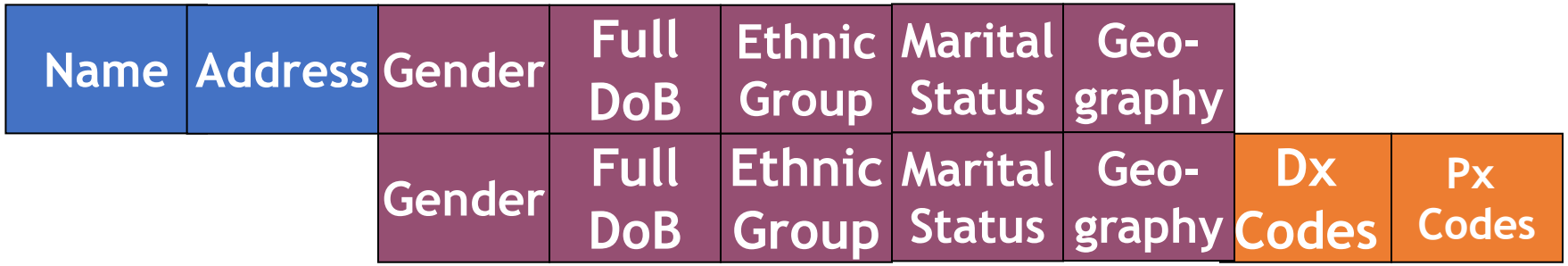
Note that this includes even fields that are not "PHI".

# *Key Resolution*

Key "*resolution*" exponentially increases with:

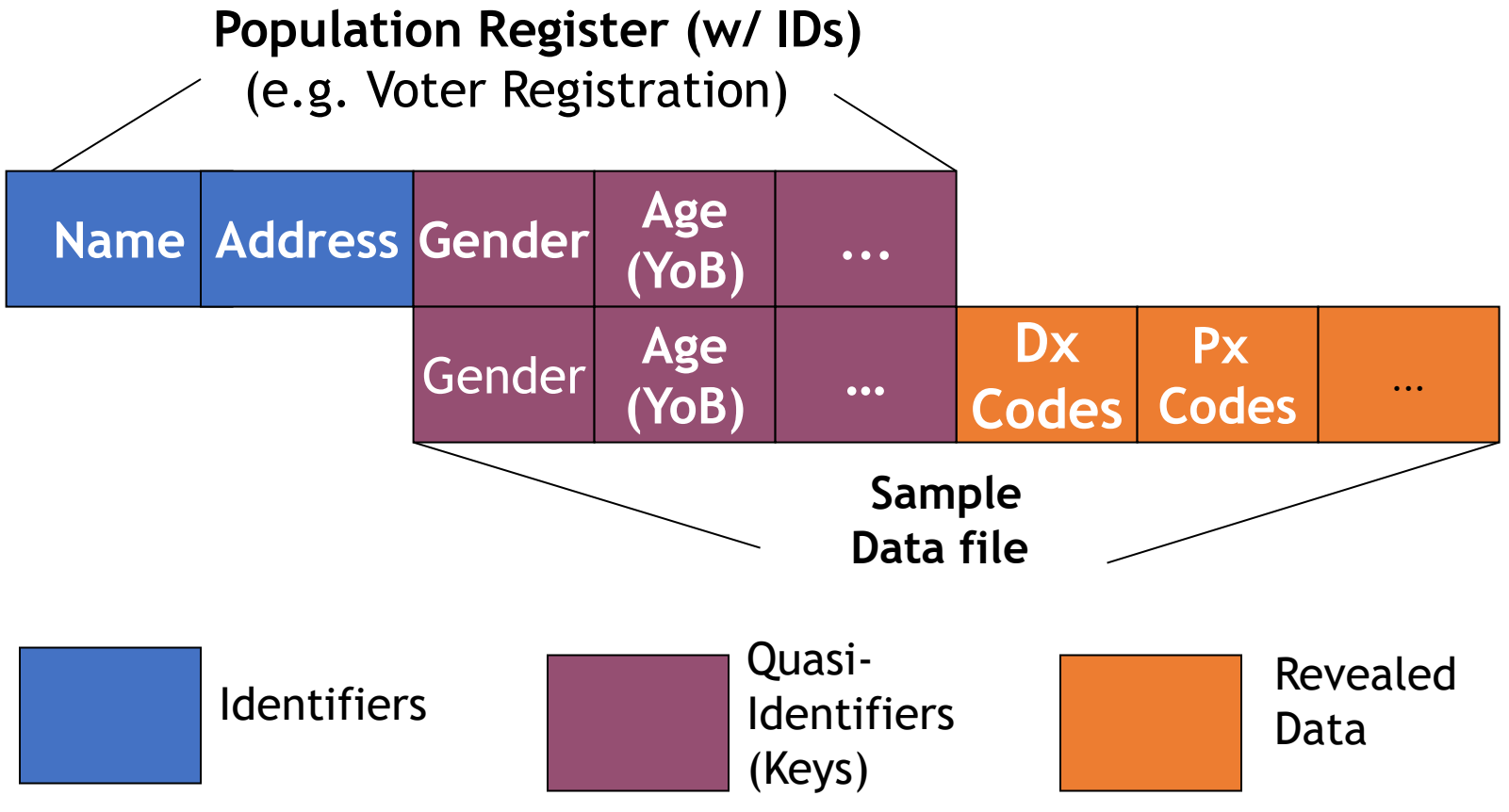
1) the number of matching fields available

2) 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.

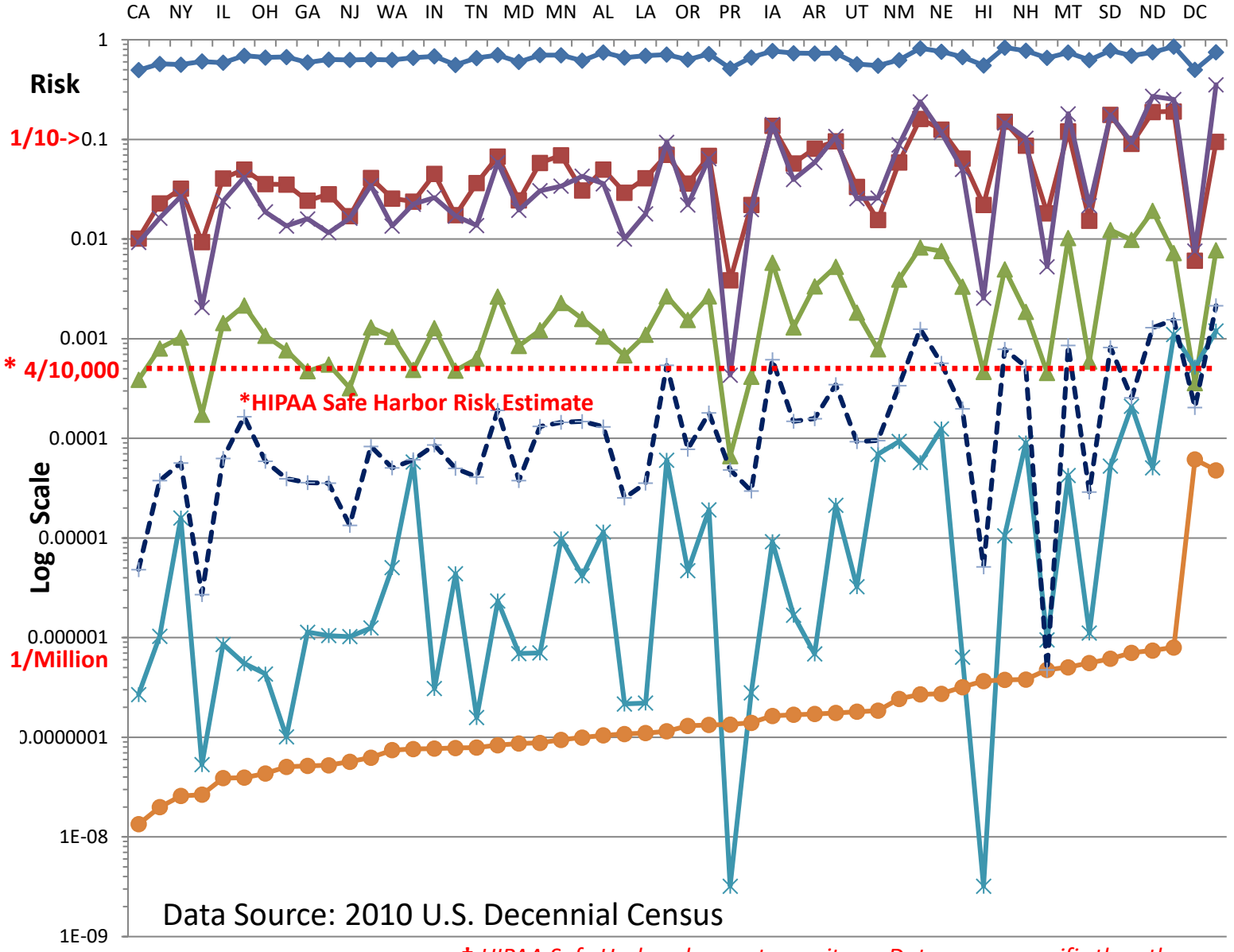


# *Sample* and *Population Uniques*

- When only one person with a particular set of characteristics exists within a given data set (typically referred to as the ***sample*** data set), such an individual is referred to as a "***Sample Unique***".
- When only one person with a particular set of characteristics exists within the entire population or within a defined area, such an individual is referred to as a "***Population Unique***".

# 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

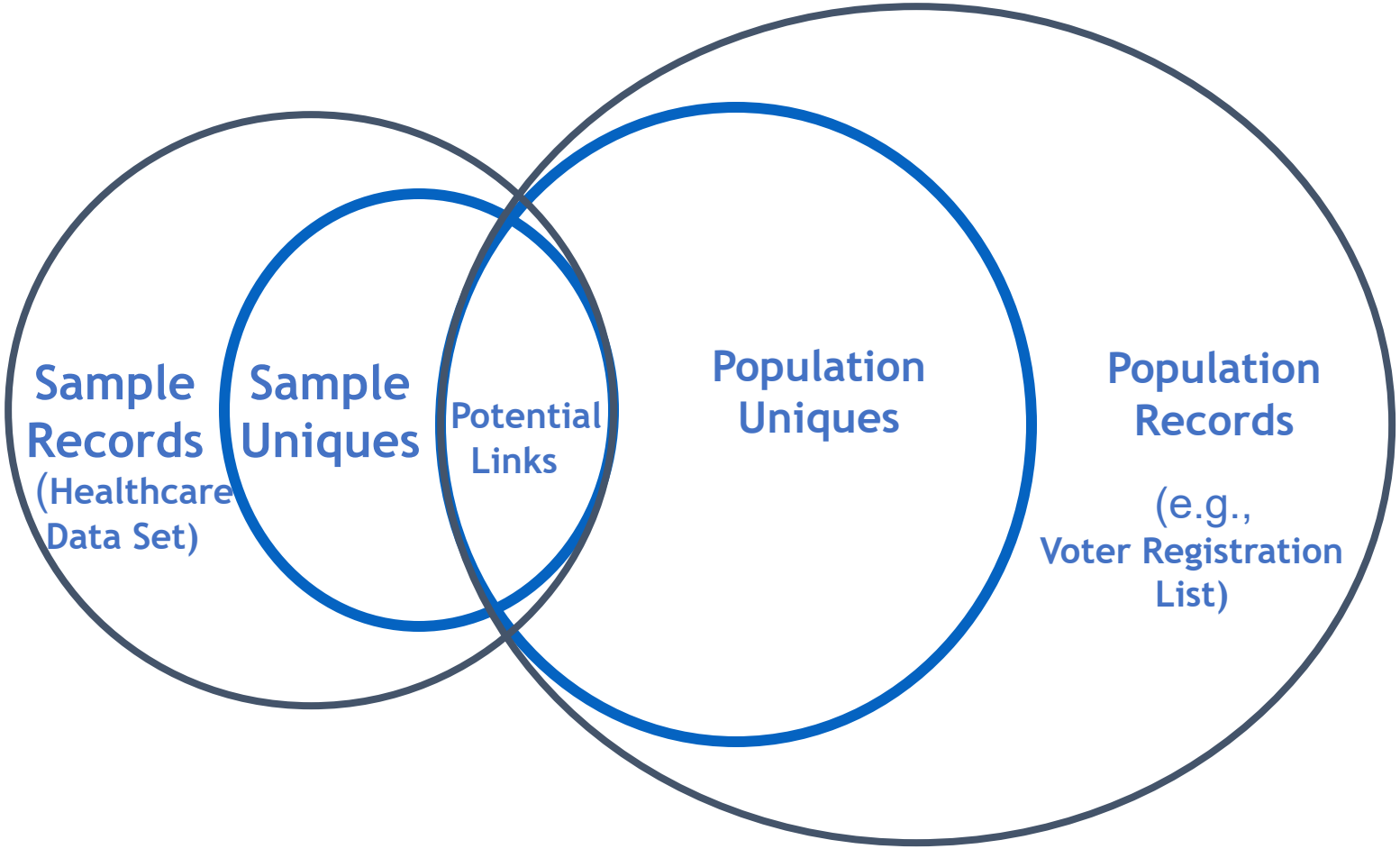
- DoB,Z5 †
- MoB,Z5
- YoB,Z5
- DoB,Z3
- MoB,Z3

- YoB,Z3 Safe Harbor
- 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).*

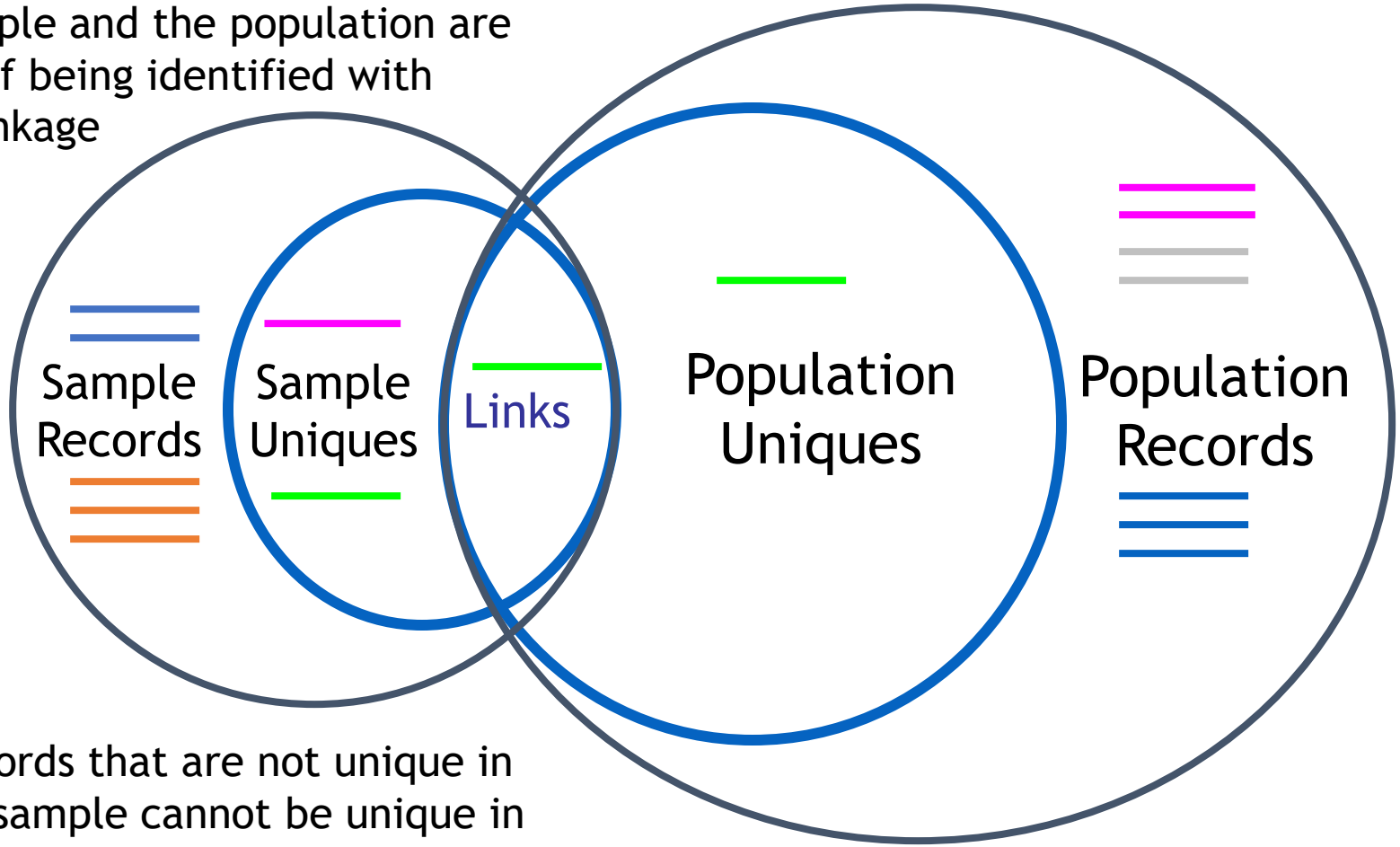
# *Measuring Disclosure Risks*



# *Linkage Risks*

Records that are unique in the sample but which aren't unique in the population, would match with more than one record in the population, and only have a probability of being identified

Only records that are unique in the sample and the population are at risk of being identified with exact linkage



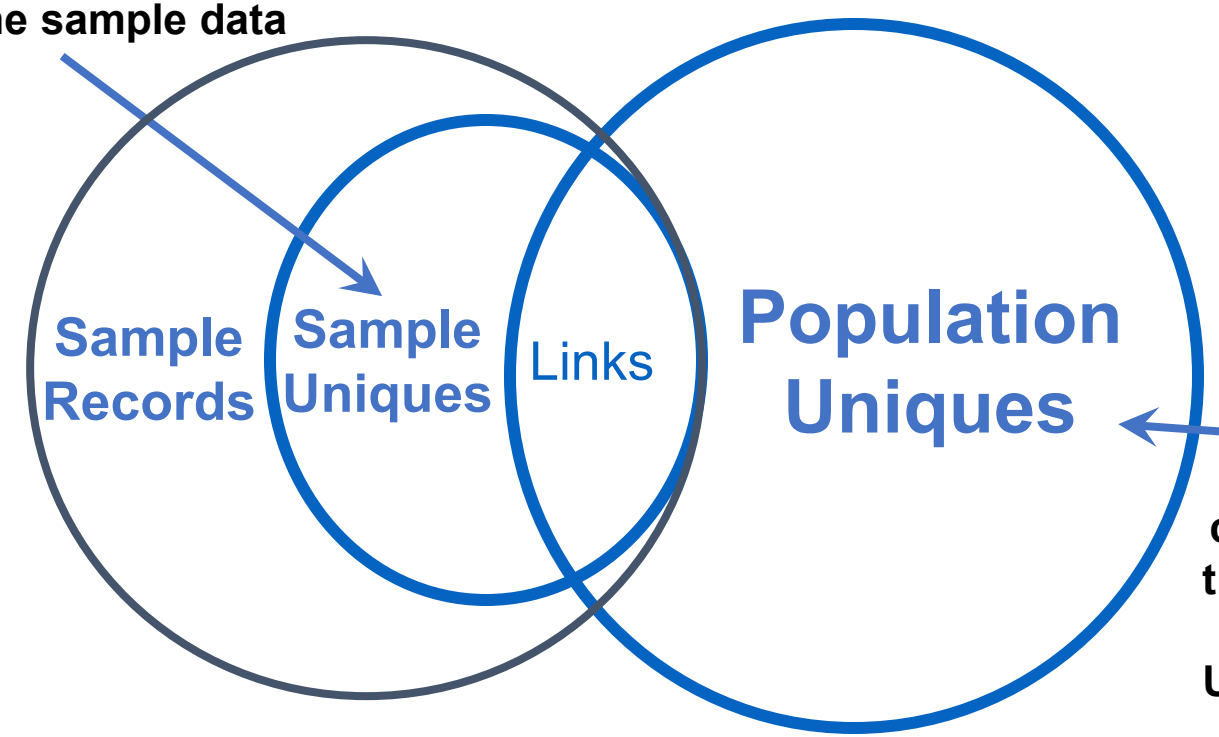

Records that are not unique in the sample cannot be unique in the population and, thus, aren't at definitive risk of being identified

Records that are not in the sample also aren't at risk of being identified

# *Estimating Disclosure Risks*

**We can determine the Sample Uniques quite easily from the sample data**

**Links / Sample Records** indicates the risk of record linkage.




**For many characteristics, the likelihood of Population Uniqueness can be estimated from statistical models of the US Census data**

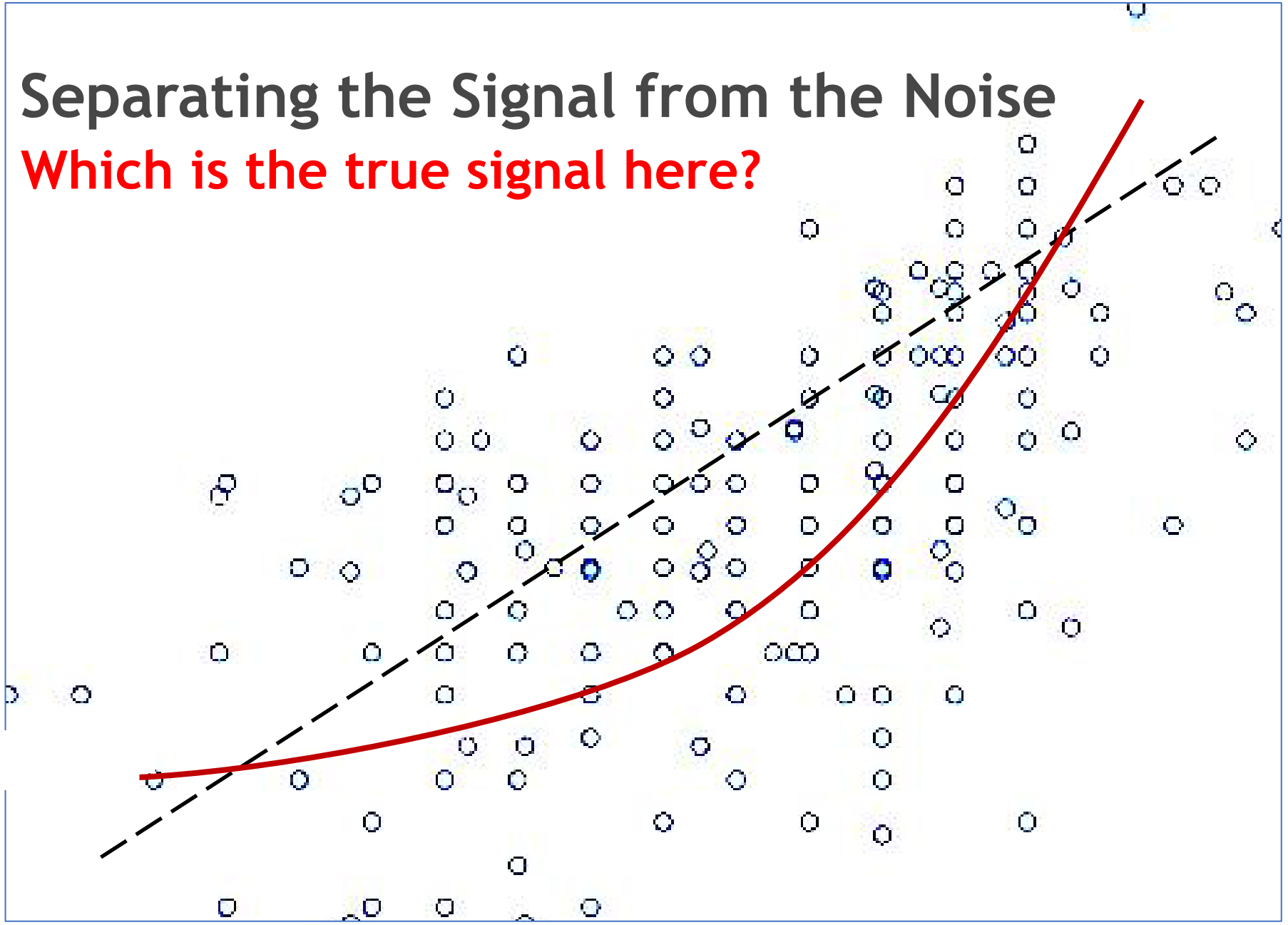
# *Balancing Disclosure Risk/Statistical Accuracy*

- Balancing disclosure risks and statistical accuracy is essential because some popular de-identification methods (e.g. k-anonymity, noise injection) can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses or data mining (distorting variance-covariance 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

# Separating the Signal from the Noise

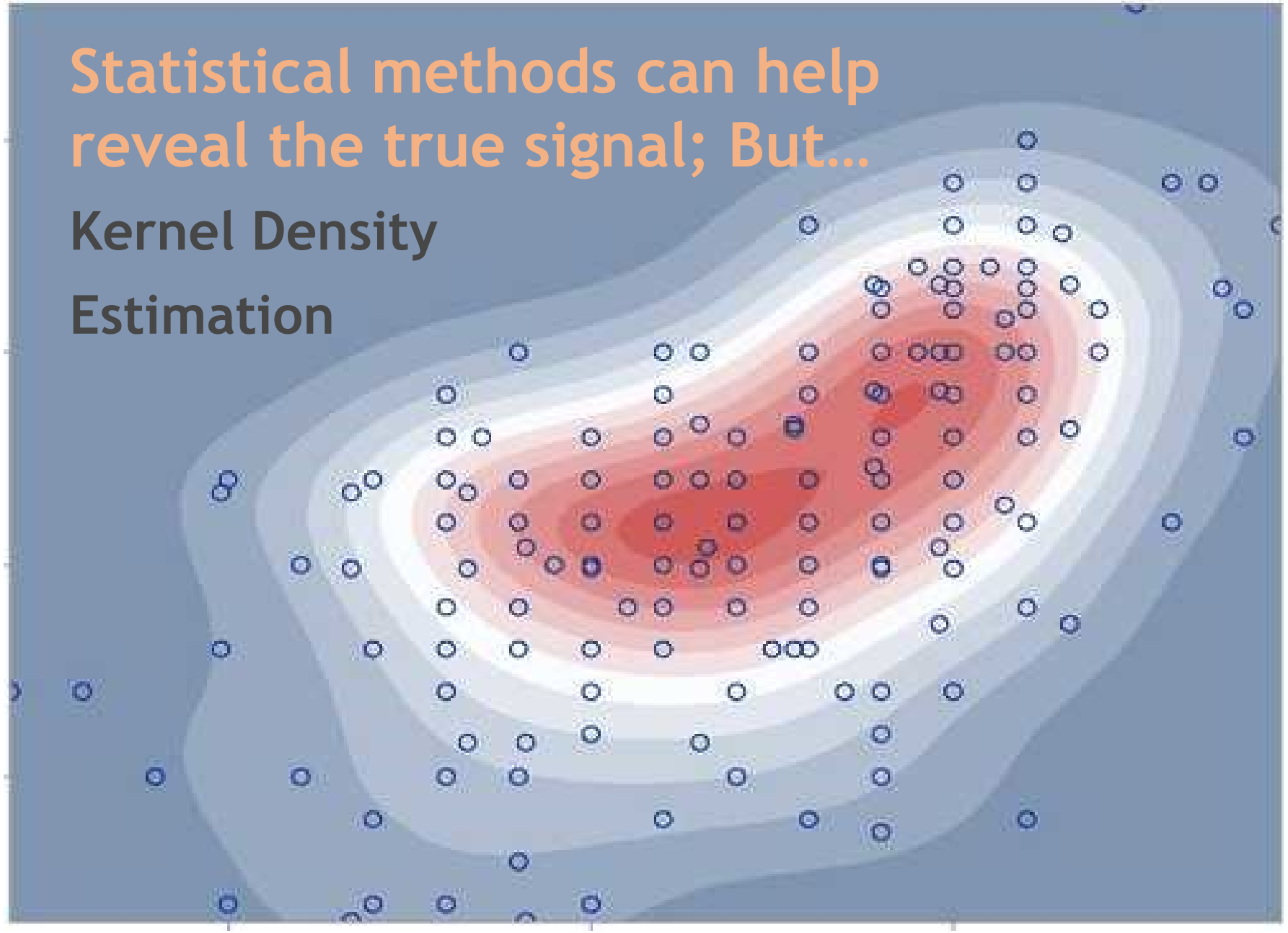
**Which is the true signal here?**



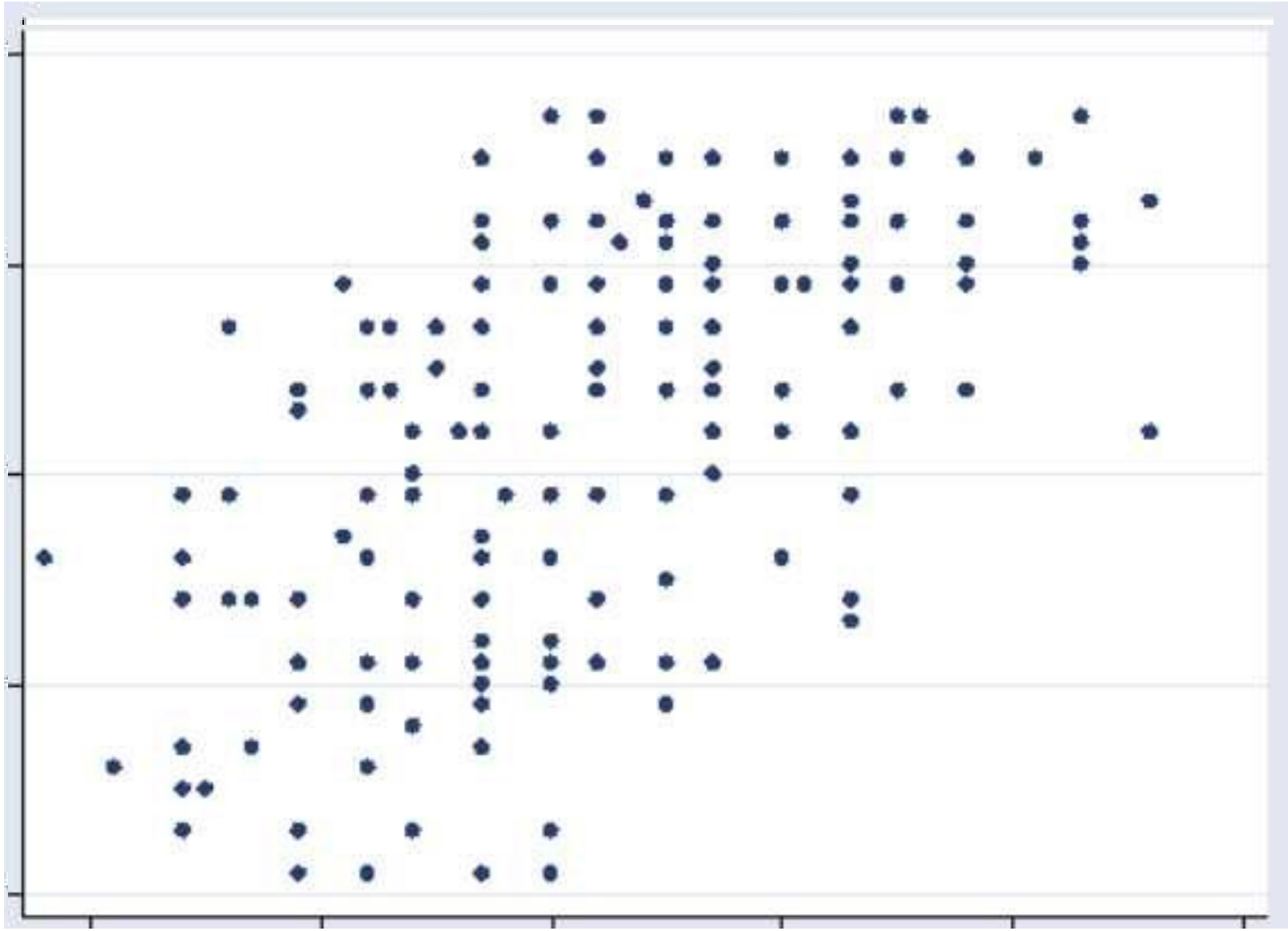
# Statistical methods can help reveal the true signal; But…

Kernel Density

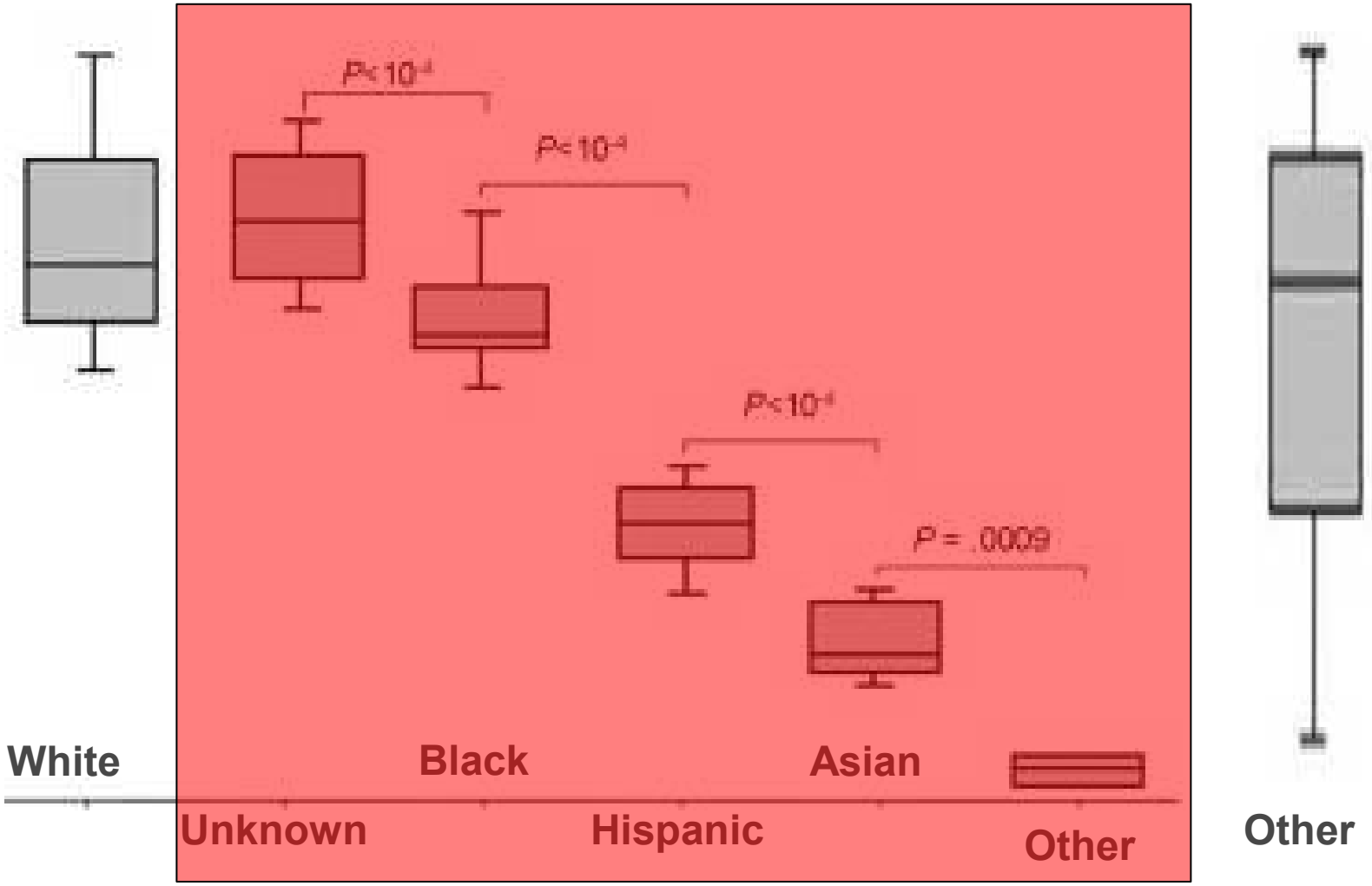
Estimation



# K-anonymity Can Distort Multivariate Relationships



# De-identification Can Hide Important Differences



Unfortunately, de-identification public policy has often been driven by largely anecdotal and limited evidence, and re-identification demonstration attacks targeted to particularly vulnerable individuals, which fail to provide reliable evidence about real world re-identification risks

# *Meanwhile, back at the "Never Ending Story"…*

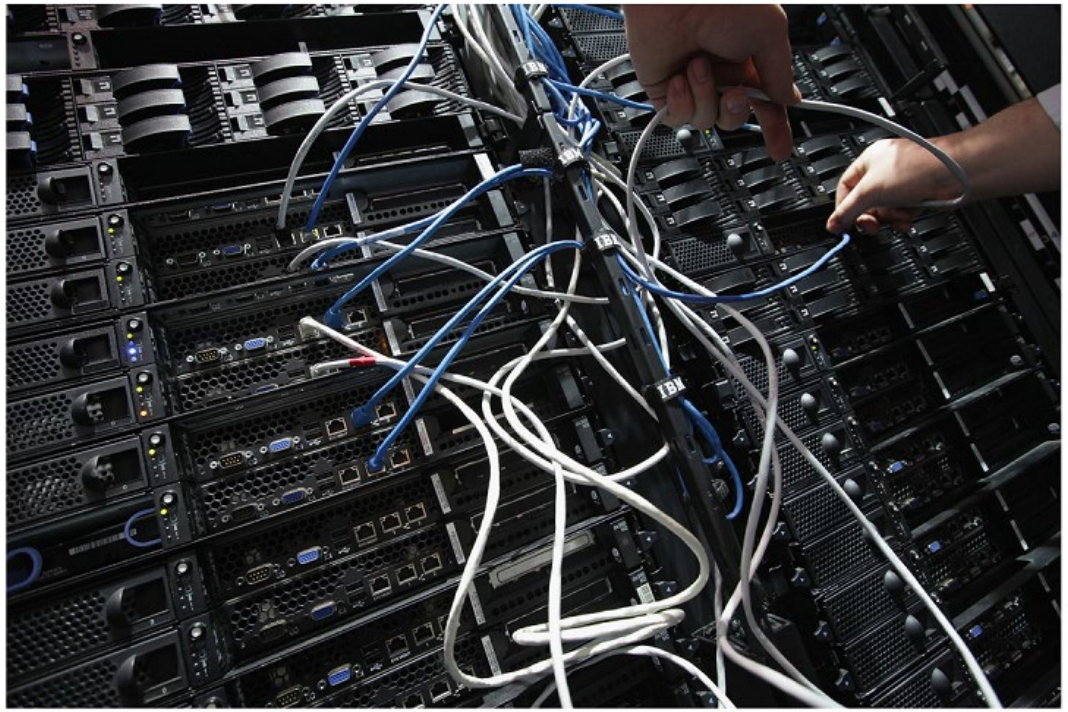
The New York Times

## *Your Data Were 'Anonymized'? These Scientists Can Still Identify You*

By Gina Kolata

July 23, 2019

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.



Scientists have found a way to identify virtually any American from any data set with just 15 attributes, like gender, ZIP code or marital status. Sean Gallup/Getty Images

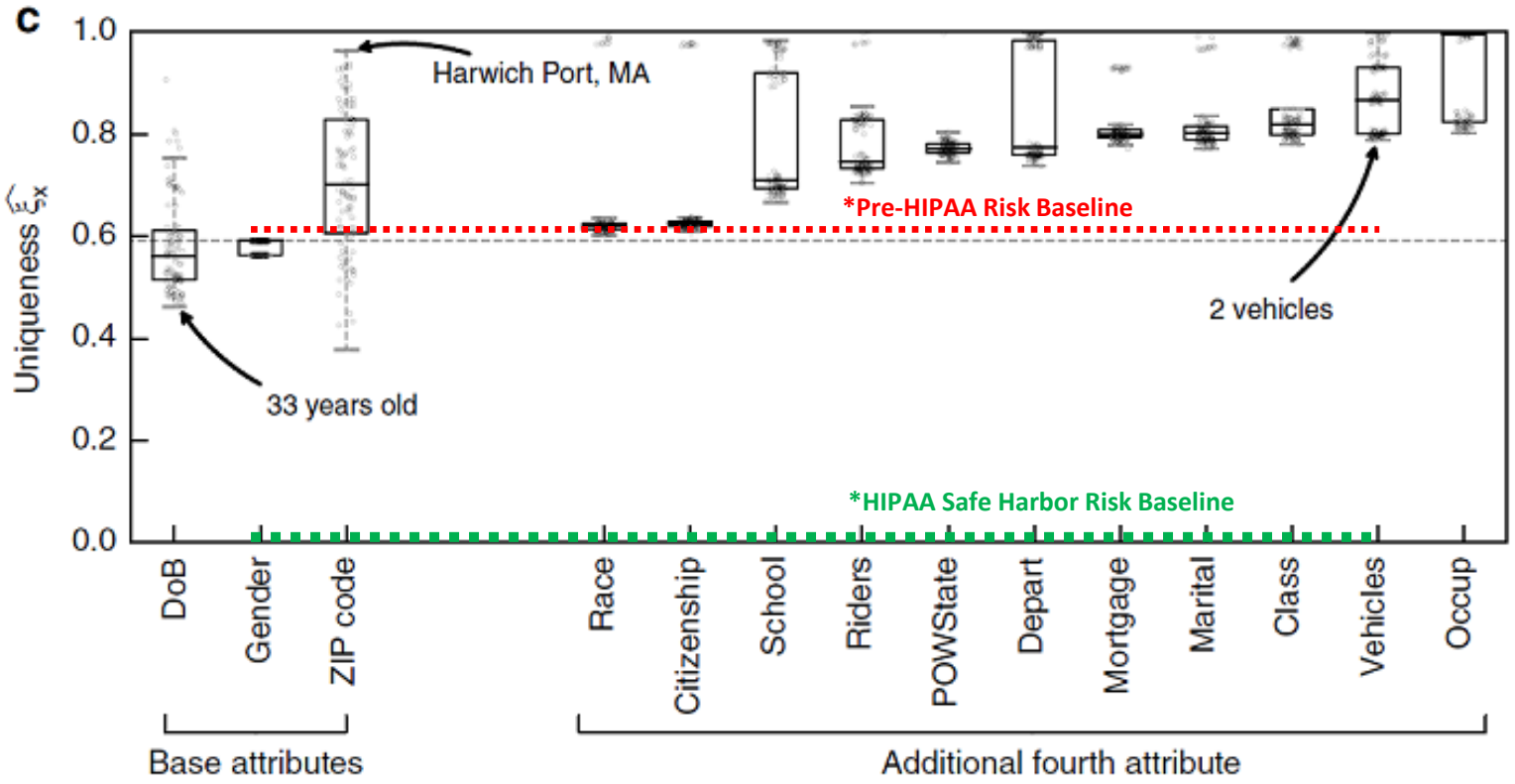
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10933-3 OPEN

# Estimating the success of re-identifications in incomplete datasets using generative models

Luc Rocher[1,2,3], Julien M. Hendrickx[1] & Yves-Alexandre de Montjoye[2,3]



# Re-identification Risks: Population Uniqueness


Data Source: 2010 U.S. Decennial Census

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

# 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

# What's the "Differential" in "Differential Privacy"?

- ***Differential Privacy*** **injects sufficient random noise into the data so for datasets with and without any of the *N* persons in the dataset, i.e., two alternative sets (one with person *i*, and without person *i*), the results will differ proportionally by a factor called Episilon (*ε*).**
- **When *ε* is small, more noise will be added to the data.**
- **So, what does this mean for differential privacy's highly-touted mathematical *Privacy "Guarantee"?***
  - *The use of the term "Guarantee" is questionable when re-identifiability depends on the selected value of Epsilon" - and the relationship between ε and re-identifiability is not easily explicated.* *(Ref: See work of Chris Clifton in References)*

# Statistical Disclosure Limitation versus Differential Privacy

- *Quasi-identifiers vs. "Everything is Personally Identifiable Information"*
- *Assumptions of Differential Privacy*
  - *All data elements are potentially knowable by data intruders and equally as useful for re-identification or attribute inference.*
  - *All data elements are equally sensitive or able to invoke privacy harms*
- *Assumptions of SDL –- Re-identification risks depend importantly on:*
    - *Replicability*
    - *Accessibility*
    - *Distinguishablity*
    - *Ability to build a comprehensive population register*

# What's to Love about Differential Privacy?

- ***Privacy Guarantees (maybe, perhaps?)***
  - **if we remember that it is only this strangely defined definition for "privacy"**
- ***Mathematical Elegance***
- ***Supports very broad assumptions about data intruder knowledge and capabilities (nearly omniscience, omnipotence and constantly co-conspiring)***
- ***Supports broad assumptions about what might be harmful in terms of data privacy attacks, impacting both re-identification risk and attribute inference.***
- ***Composability*** **(computer scientists/mathematicians love this!)**
- ***Enforces Consistency in your "privacy budget".***

# What's Not to Love about Differential Privacy?

- *Neither the terms "Privacy" or "Guarantee" mean what most people think they mean…*
- *The complexity of communicating what it does and how it does it to the public*
- *The "accuracy costs" that are incurred by its very broad assumptions*
- *Ethical dilemmas posed by the transfer of these accuracy costs to data for other purposes and individuals*
- *Differential Privacy strictly enforces the "privacy", but only optionally enforces the accuracy issues through a wise, reasoned and empirically analyzed and justifiable selection of epsilon.*
- *It is not without completely free of potential avenues of attack*
  - *Repeated instantiations can be revelatory*
  - *Correlated observations don't receive the same guarantees*

# *References for Differential Privacy*


1. Clifton, C.; Tassa, T. On Syntactic Anonymity and Differential Privacy. Transactions On Data Privacy 6 (2013) 161–183
2. Lee J., Clifton C. How Much Is Enough? Choosing ε for Differential Privacy. In: Lai X., Zhou J., Li H. (eds) Information Security. ISC 2011. Lecture Notes in Computer Science, vol 7001. Springer, Berlin, Heidelberg
3. Lee, J.; Clifton C. Differential Identifiability. KDD ’12, August 12–16, 2012, Beijing, China.

# Khaled El Emam. PhD

## Professor, University of Ottawa
## SVP and General Manager, Replica Analytics

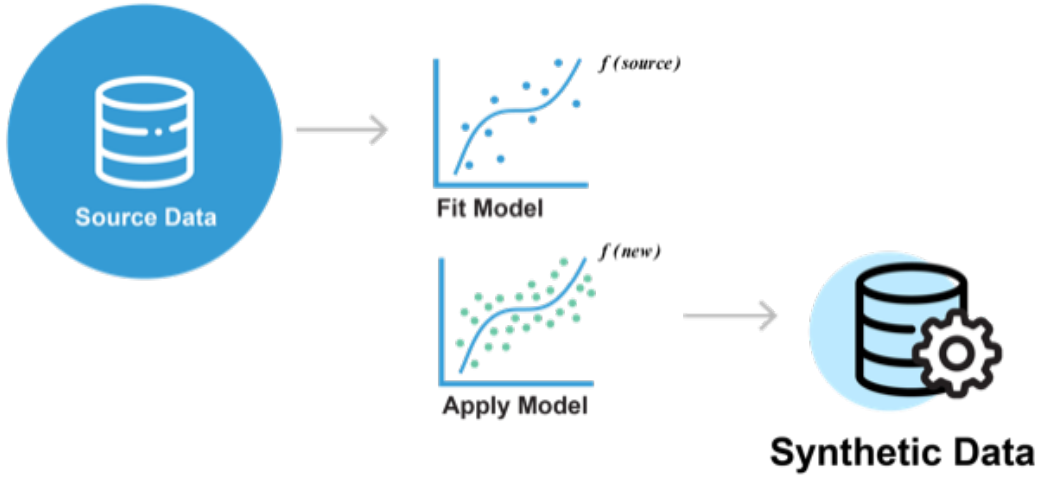
Dr. Khaled El Emam is co-founder and General Manager of Replica Analytics, a company that develops data synthesis technology. He is the Canada Research Chair (Tier 1) in Medical AI at the University of Ottawa, where he is a Professor in the School of Epidemiology and Public Health.  He is also a Senior Scientist at the Children's Hospital of Eastern Ontario Research Institute.

As an entrepreneur, Khaled founded or co-founded six companies involved with data management and data analytics. In 2003 and 2004, he was ranked as the top systems and software engineering scholar worldwide by the Journal of Systems and Software based on his research on measurement and quality evaluation and improvement.

Previously, Khaled was a Senior Research Officer at the National Research Council of Canada. He also served as the head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany.

Khaled held the Canada Research Chair in Electronic Health Information at the University of Ottawa from 2005 to 2015. He has a PhD from the Department of Electrical and Electronics Engineering, King's College, at the University of London, England.

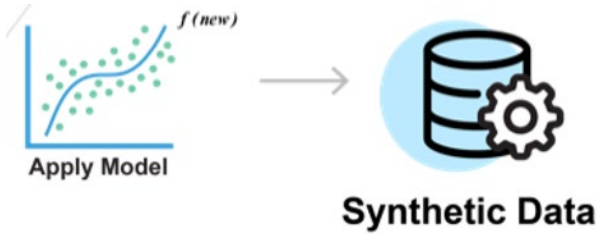
# The Synthesis Process



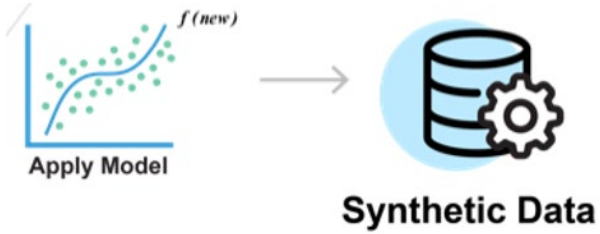

| COU1A | AGECAT | AGELE70 | WHITE | MALE | BMI |
|---|---|---|---|---|---|
| United States | 2 | 1 | 1 | 1 | 33.75155 |
| United States | 2 | 1 | 1 | 0 | 39.24707 |
| United States | 1 | 1 | 1 | 0 | 26.5625 |
| United States | 4 | 1 | 1 | 1 | 40.58273 |
| United States | 5 | 0 | 0 | 1 | 24.42046 |
| United States | 5 | 0 | 1 | 0 | 19.07124 |
| United States | 3 | 1 | 1 | 1 | 26.04938 |
| United States | 4 | 1 | 1 | 1 | 25.46939 |

K. El Emam, L. Mosquera, C. Zheng: "Optimizing the Synthesis of Clinical Trial Data Using Sequential Trees", Journal of the American Medical Informatics Association 28(1): 3-13, 2021.

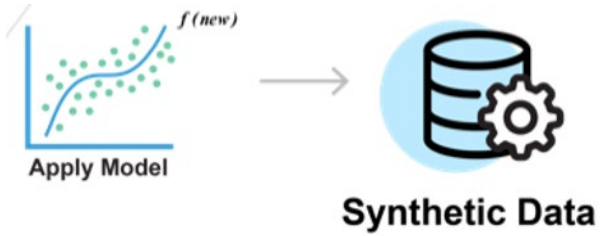
# A simulator exchange allows data to be made available without sharing actual data




Synthetic Data

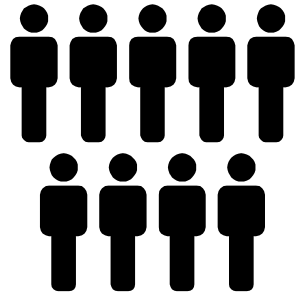



Synthetic Data




Synthetic Data

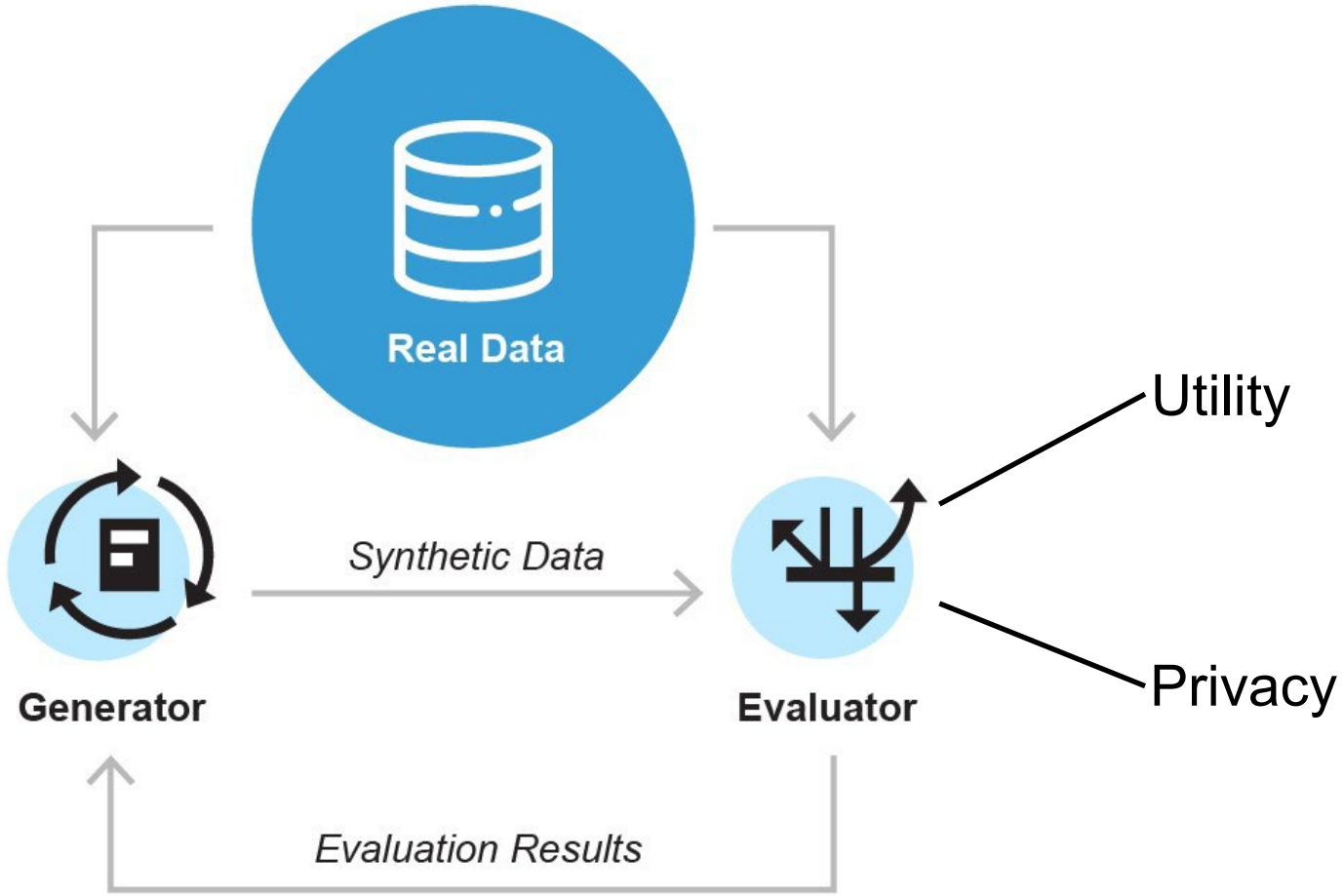
## Data Consumers



**Additional Clarifications**

- The simulators would not be given to the data consumers – they would only have access to them through an interface.
- This access would be monitored and throttled to reduce the risk of attacks on the models.
- Data consumers would also need to agree to terms of use around the access to the simulators.

# Training a generative model uses a utility – privacy loss function



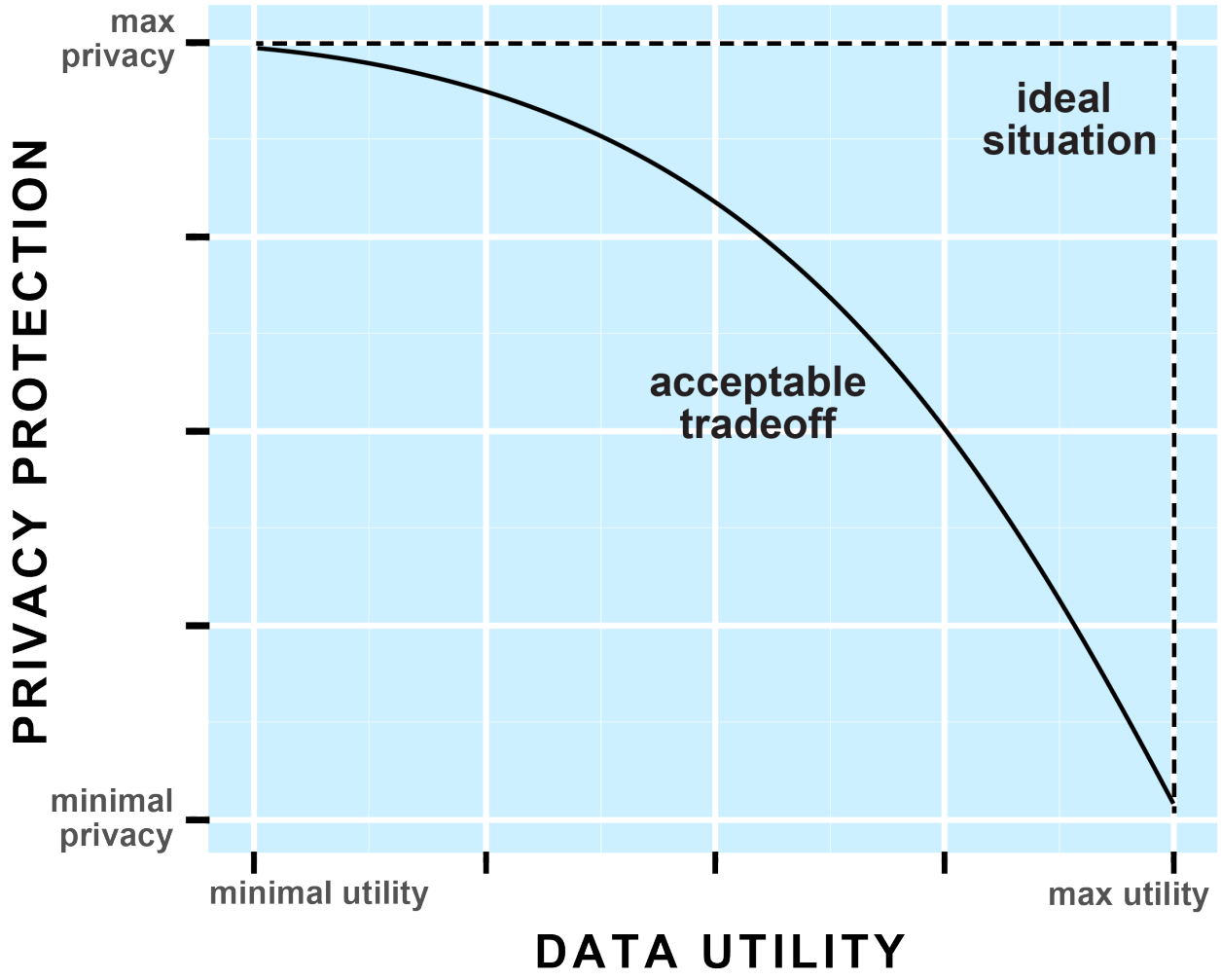
# There are seven common use cases for synthetic data

1. Machine learning

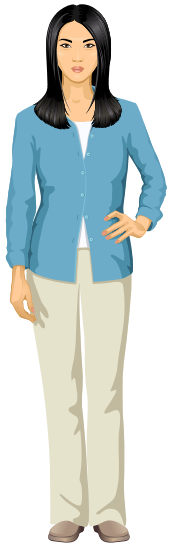
   *(model evaluation, data augmentation, sharing ML models)*
2. Software testing
3. Education, training, and hackathons
4. Data retention
5. Vendor assessment
6. Internal secondary use

   *(exploratory and detailed analytics)*
7. External data sharing

# Privacy-Utility Trade-off



# Attribution disclosure: find a similar record in the synthetic data and learn something new



Quasi-identifiers

New Information

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1975 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1972 | 0074-5182 |
| Female | 1993 | 0078-0379 |
| **Female** | **1989** | **65862-403** |
| Male | 1991 | 55714-4446 |
| Male | 1992 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1971 | 55289-324 |
| Female | 1996 | 54868-6348 |
| Male | 1980 | 53808-0540 |

K. El Emam, L. Mosquera, J. Bass: "Evaluating Identity Disclosure Risk in Fully Synthetic Health Data", Journal of Medical Internet Research, 22(11), 2020.

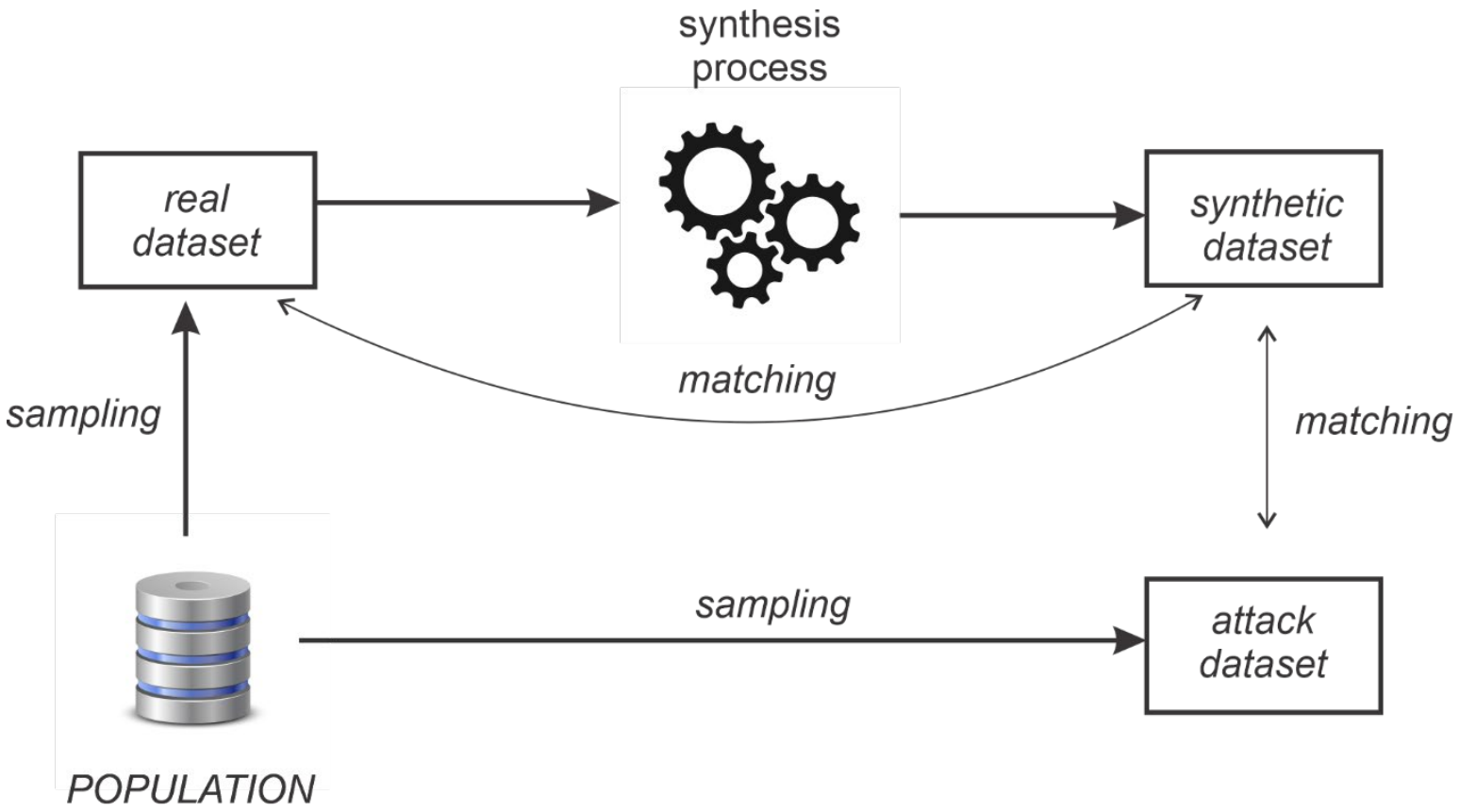
# Evaluations of attribution risks show that it is low in multiple studies across multiple datasets

| Dataset | Fully Synthetic Data | Original Data |
|---|---|---|
| **Washington Hospital Data (Discharge)** | 0.0197 | 0.098 |
| **Canadian COVID-19 Data (Public Health)** | 0.0086 | 0.034 |

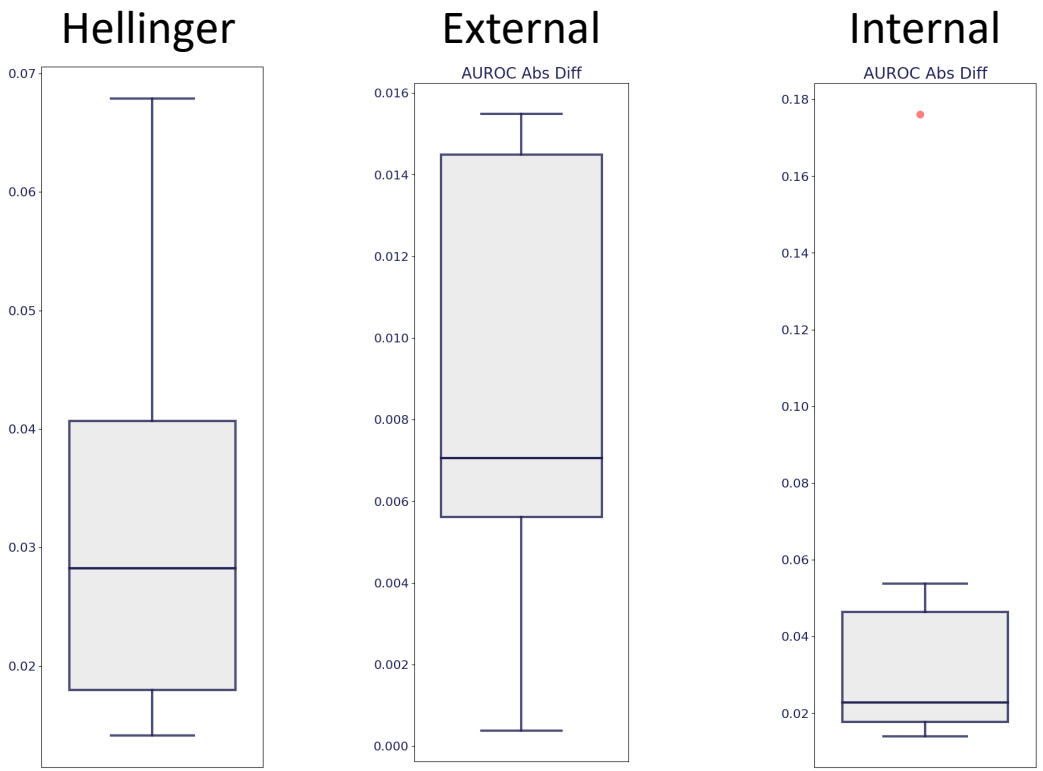
A commonly used risk threshold = 0.09

K. El Emam, L. Mosquera, J. Bass: "Evaluating Identity Disclosure Risk in Fully Synthetic Health Data", Journal of Medical Internet Research, 22(11), 2020.

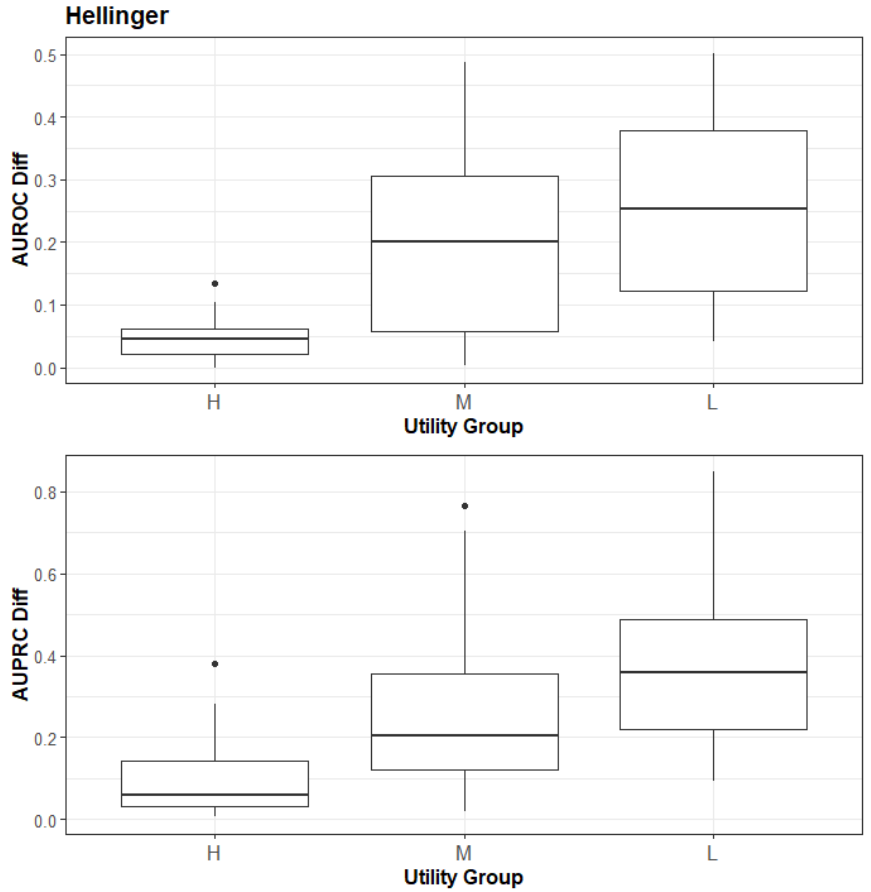
# Membership disclosure



# Example generic utility metrics for individual synthetic datasets

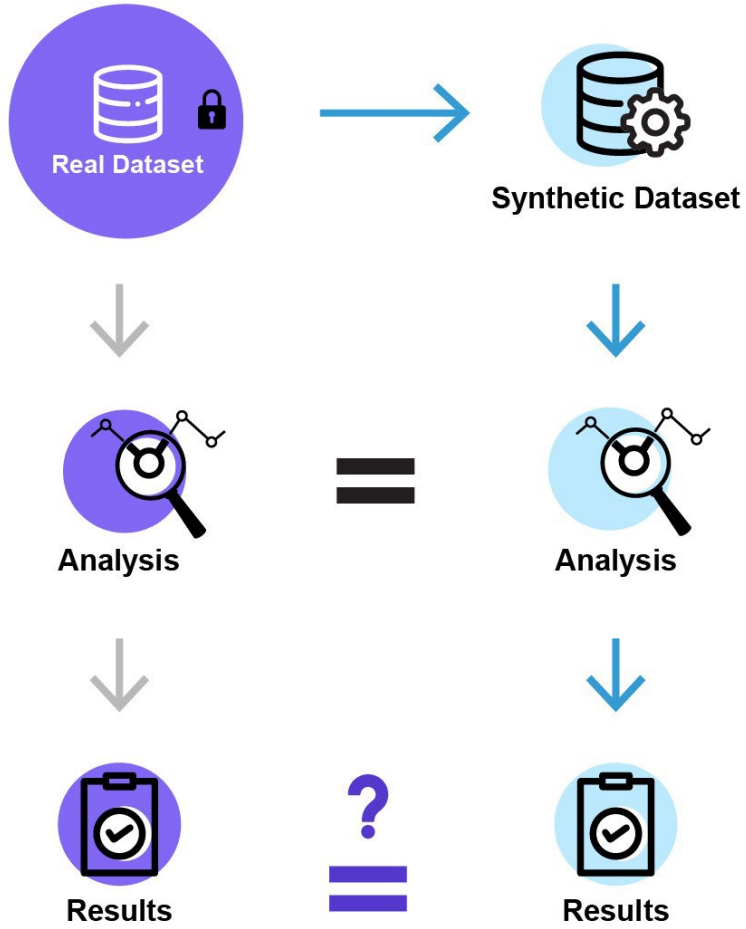


# Broad utility metrics can rank SDG methods by their workload performance

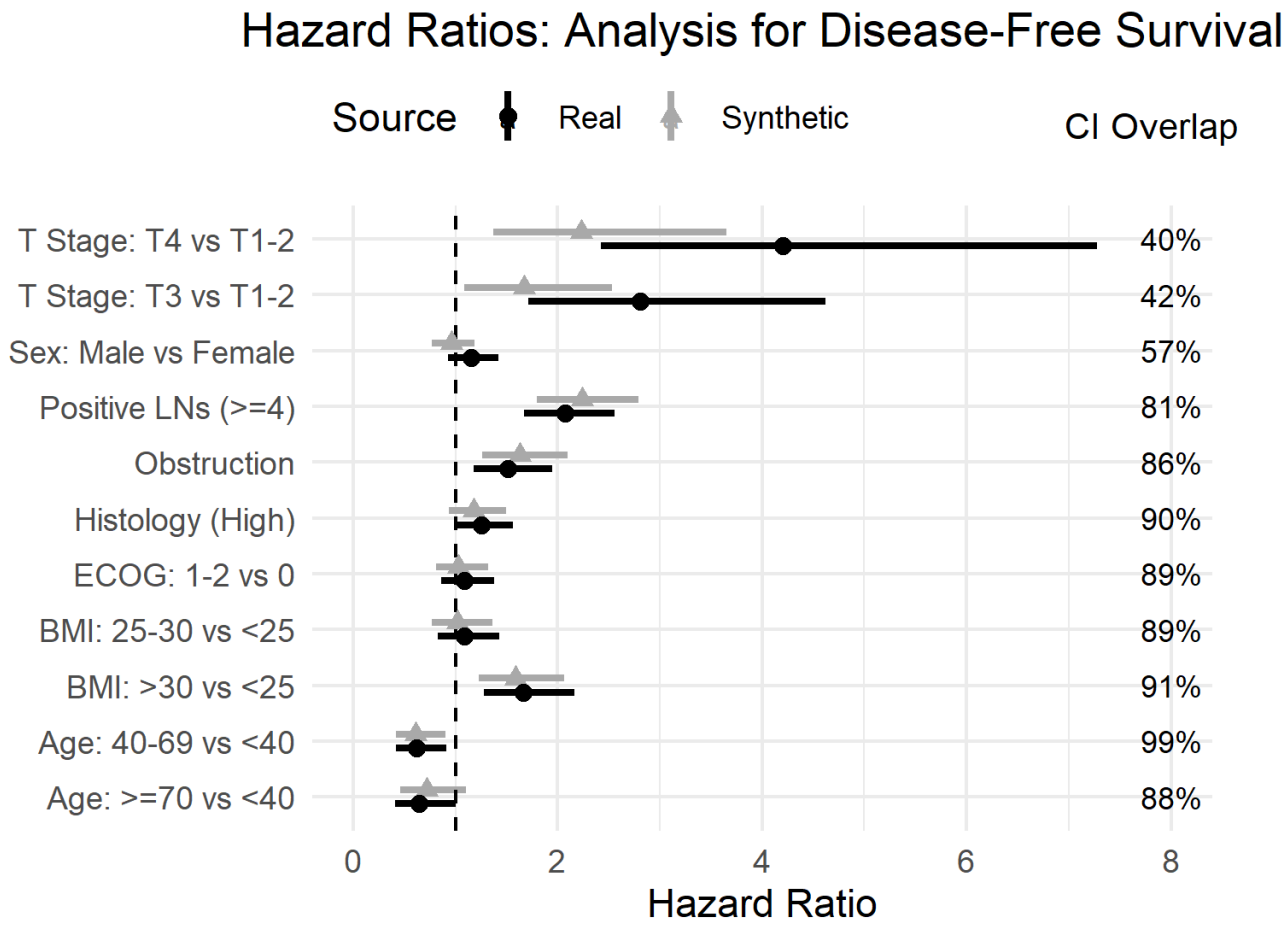



K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna: "Utility Metrics for Evaluating Synthetic Health Data Generation Methods: A Validation Study", JMIR Medical Informatics (in press), 2022.

# To evaluate utility one can compare the analysis results from real and synthetic data



# Comparing real and synthetic data: Adjusted model of impact of bowel obstruction on DFS




Z. Azizi, M. Zheng, L. Mosquera, L. Pilote, K. El Emam: "Can synthetic data be a proxy for real clinical trial data? A validation study", BMJ Open, 11:e043497, 2021.

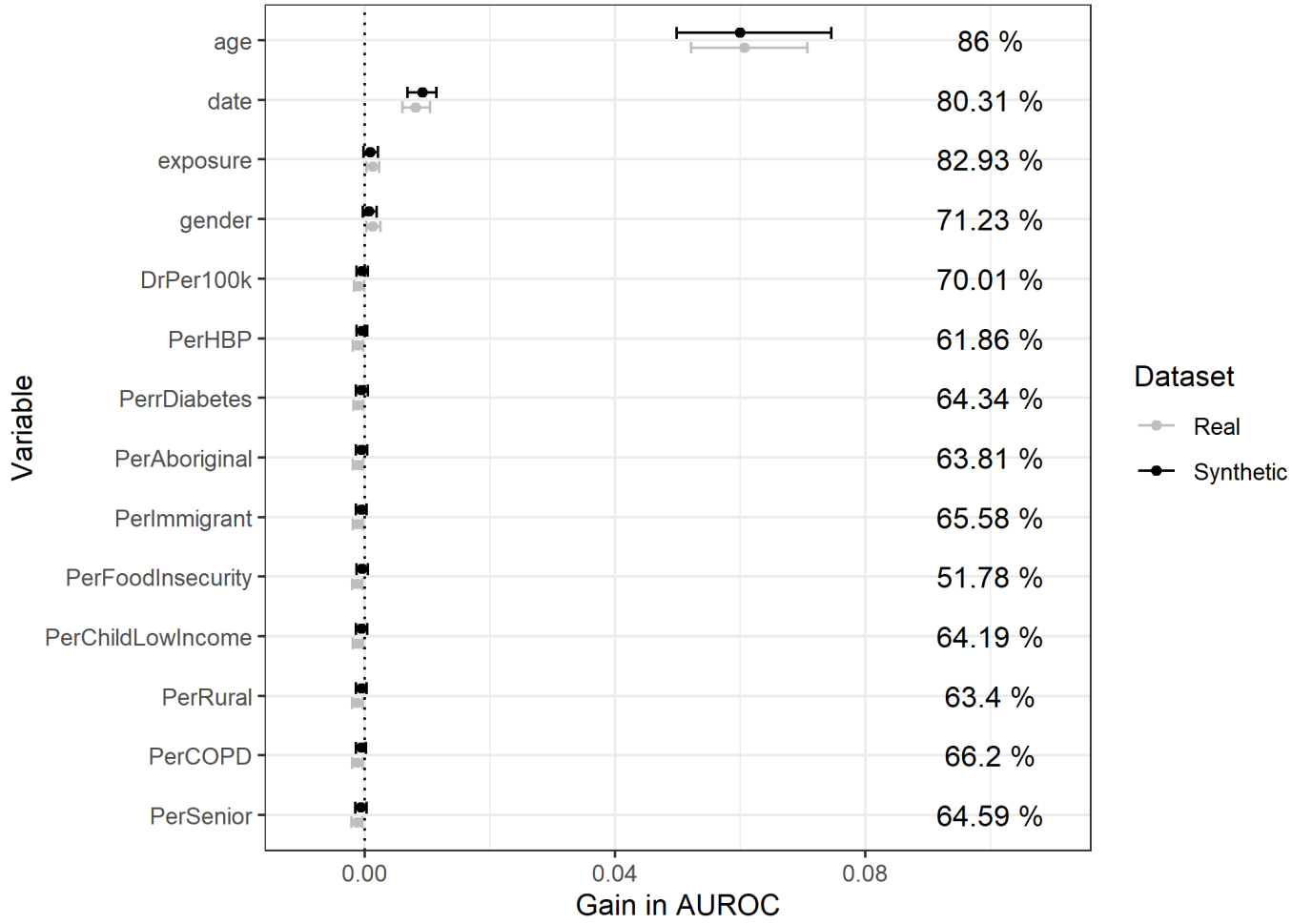

6

# Model accuracy for predicting COVID-19 mortality in Ontario

| Accuracy metric | Real data | Synthetic data | CI overlap |
|---|---|---|---|
| **AUROC** | 0.945 (0.941–0.948) | 0.940 (0.936–0.945) | 45.50% |
| **AUPRC** | 0.340 (0.314–0.368) | 0.313 (0.286–0.342) | 52.02% |

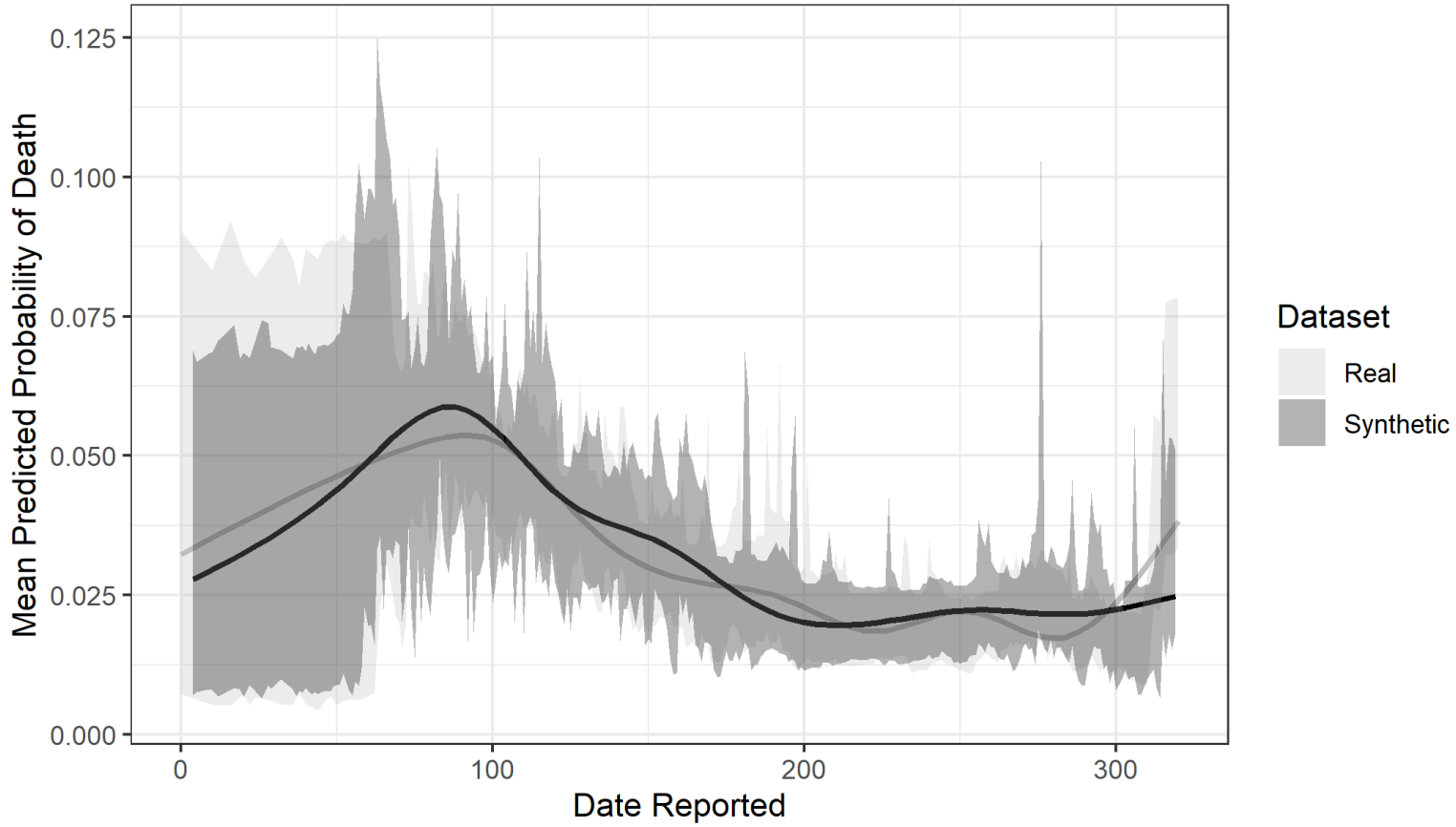
K. El Emam, L. Mosquera, E. Jonker, H. Sood: "Evaluating the Utility of Synthetic COVID-19 Case Data", JAMIA Open, 14(1):ooab012, January 2021.

# Variable Importance



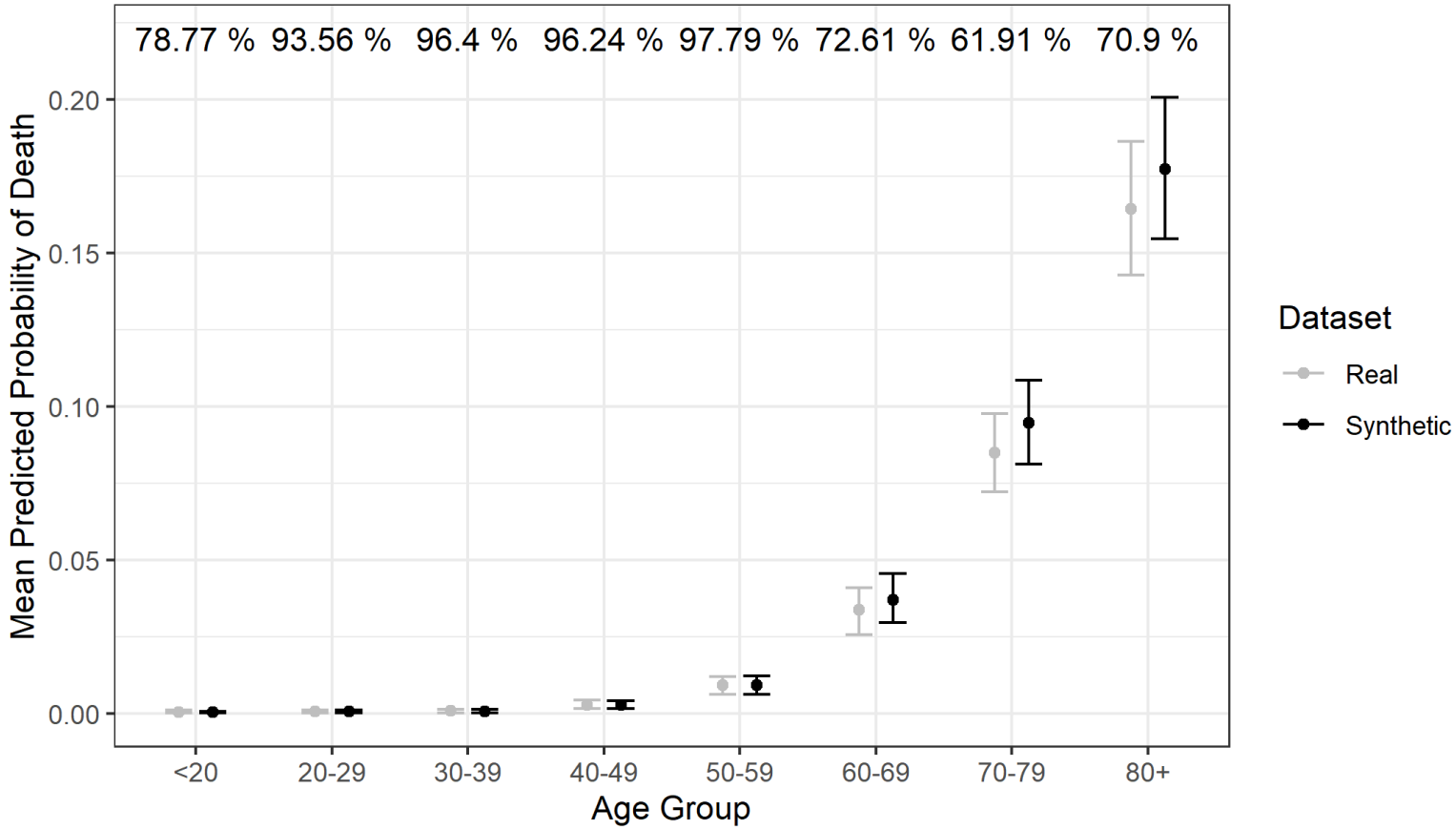

K. El Emam, L. Mosquera, E. Jonker, H. Sood: "Evaluating the Utility of Synthetic COVID-19 Case Data", JAMIA Open, 14(1):ooab012, January 2021.

# Mortality Over Time




K. El Emam, L. Mosquera, E. Jonker, H. Sood: "Evaluating the Utility of Synthetic COVID-19 Case Data", JAMIA Open, 14(1):ooab012, January 2021.

# Mortality By Age




K. El Emam, L. Mosquera, E. Jonker, H. Sood: "Evaluating the Utility of Synthetic COVID-19 Case Data", JAMIA Open, 14(1):ooab012, January 2021.

# Questions + Contact

## Ann Waldo, JD

**Waldo Law Offices**

## Allison Bender, JD

**Denton's**

## Daniel Barth-Jones, PhD

**Assistant Professor of Clinical Epidemiology**
**Mailman School of Public Health, Columbia University**

## Khaled El Emam, PhD

**SVP and General Manager, Replica Analytics**
**Professor, University of Ottawa**

# Reserve Slides for Questions

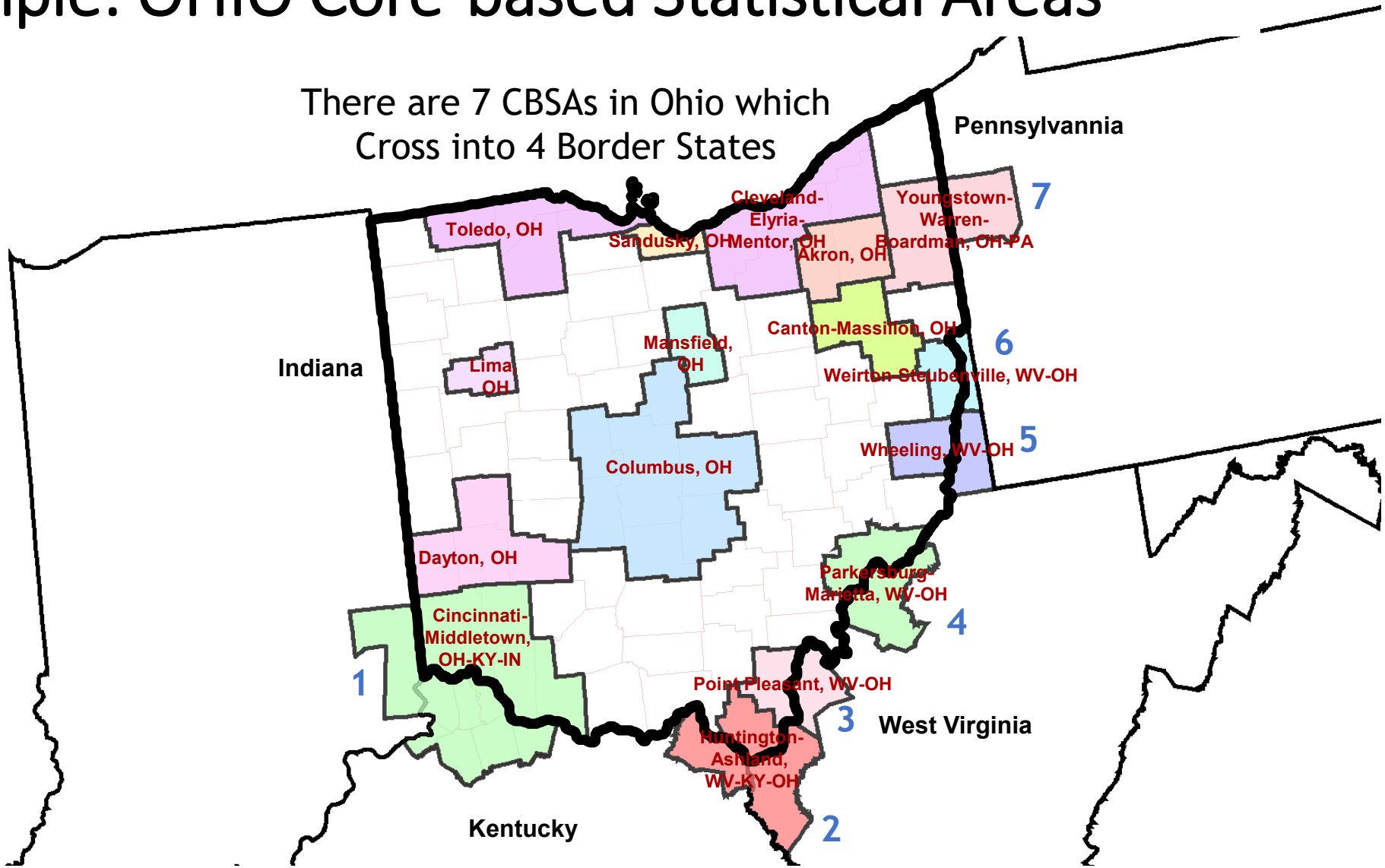
# HIPAA §164.514(b)(2)(i) -18 “Safe Harbor” Exclusions

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)

# Example: OHIO Core-based Statistical Areas



# *Data Privacy Concerns are Far Too Important (and Complex) to be summed up with Catch Phrases or "Anecdata"*

Eye-catching headlines and twitter-buzz announcing "***There's No Such Thing as Anonymous Data***" might draw the public's attention to broader and important concerns about data privacy in this era of "Big Data",

but such statements are essentially meaningless, even misleading, for further generalization without consideration of the specific de/re-identification contexts -- including the precise data details (e.g., number of variables, resolution of their coding schemas, special data properties, such as spatial/geographic detail, network properties, etc.) de-identification methods applied, and associated experimental design for re-identification attack demonstrations.

**Good Public Policy demands reliable scientific evidence…**

# Re-identification Demonstration Attack Summary

| Re-identification Attacks | Quasi-Identifers (w/ HIPAA Safe Harbor exclusion data in Red) | Vulnerable Subgroup Targeted? | Used Stat. Sampling | Individuals w/ Alleged/Verified Re-identification | At-Risk Sample Size | Notable Headlines & Quotes | Attack Against HIPAA Compliant (or SDL Protected) Data? | Demonstrated Re-identification Risk |
|---|---|---|---|---|---|---|---|---|
| Governor Weld [1,2] | Zip5, Gender, DoB | Yes | No | n=1 | 99,500 | *"Anonymized" Data Really Isn't* [27] | No | 0.00001 |
| AOL [3] | Free Text from Search Queries w/ Name, Location, etc | Yes | No | n=1 | 657,000 | *A Face is Exposed* [3] | No | 0.0000015 |
| Netflix [4] | Movie Ratings & Dates | Yes | No | n=2 | 500,000 | *"...successfully identified 99% of people in Netflix database"* [28] | No | 0.000004 |
| ONC Safe Harbor [5] | Zip3, YoB, Gender, Marital Status, Hispanic Ethnicity | No | N/A | n=2 | 15,000 | [ Press Did Not Cover This Study ] | Yes | 0.00013 |
| Heritage Health Prize [6,7,8,9] | Age, Sex, Days in Hospital, Physician Specialty, Place of Service, CPT Code, Days Since First Claim, ICD-9 Diagnosis | Yes | No | n=0 | 113,000 | *To best of my judgment, reidentification is within realm of possibility* [8] *El Emam estimated < 1% of Pts could be re-identified. Narayanan estimated > 12% of Pts were identifiable.* [29] | Yes | 0.0 |
| Y-Chromosome STR Surname Inference [10,11] - Simulation Study Part | Y-STR DNA Sequences* Age in Years & State | No | N/A, Simulation | Not Attempted: Simulated Results | ~150 Million US Males | *"nice example of how simple it is to re-identify de-identified samples"* [30] | *No? (Safe Harbor vs. Expert Determination) *Safe Harbor Excludes: *Any unique identifying #, characteristic or code* | .12 (For Males Only), after accounting for 30% False Positive Rate |
| - CEU Attack Part | Age, Utah State, Genealogy Pedigrees & Mormon Ancestry | Yes, Highly Targeted | No | n=5 w/ Y-STR Alone, (but w/ Geneology Amplification n=50) | ? | *DNA Hack Could Make Medical Privacy Impossible* [31] | | Not Clearly Calculable for CEU Attack |
| Personal Genome Project [12,13,14] | Zip5, Gender, DoB | No | N/A | n=161 | 579 | *"...re-identified names of > 40% anonymous participants"* [32] *re-identified 84 to 97% of sample of PGP volunteers* [33] | No | 0.28 (w/ Embedded Names Excluded) |
| Washington St. Hospital Discharge [15,16] | Hospital Data w/ Diagnoses, Zip5, Month/Yr of Discharge | Yes | No | n=40 (8 verified) from 81 News Reports | 648,384 | *"...how new stories about hospital visits in Washington State leads to identifying matching health record 43% of the time"* [34] | No | 0.000062 |
| Cell Phone "Unicity" [17] | High Resolution Time (Hours) and Cell Tower Location | No | N/A | Not Attempted | 1.5 Million | *"four spatio-temporal points enough to uniquely identify 95%"* [17] | No | 0.0 |
| NYC Taxi [18,19] | High Resolution Time (Minutes) and GPS Locations | Yes | No | n=11 | 173 Million Rides | *How Big Brother Watches You With Metadata* [35] | No | 0.0000001 |
| Credit Card "Unicity" [20,21,22,23,24,25,26] | High Resolution Time (Days), Location and Approx. Price | No | N/A | Not Attempted | 1.1 Million | *With a Few Bits of Data, Researchers Identify 'Anonymous' People* [36] | No | 0.0 |

- *Publicized attacks are on data without HIPAA/SDL de-identification protection.*
- *Many attacks targeted especially vulnerable subgroups and did not use sampling to assure representative results.*
- *Press reporting often portrays re-identification as broadly achievable, when there isn't any reliable evidence supporting this portrayal.*

# Re-identification Demonstration Attack Summary

- For Ohm's famous "Broken Promises" attacks (Weld, AOL, Netflix) a total of n=4 people were re-identified out of 1.25 million.
- For attacks against HIPAA de-identified data (ONC, Heritage*), a total of n=2 people were re-identified out of 128 thousand.
  - ONC Attack Quasi-identifers: Zip3, YoB, Gender, Marital Status, Hispanic Ethnicity
  - Heritage Attack Quasi-identifiers*: Age, Sex, Days in Hospital, Physician Specialty, Place of Service, CPT Procedure Codes, Days Since First Claim, ICD-9 Diagnoses (*not complete list of data available for adversary attack)
  - Both were "adversarial" attacks.
- For all attacks listed, a total of n=268 were re-identified out of 327 million opportunities.

***Let's get some perspective on this…***

Obviously, This slide is **BLACK**

So clearly, De-identification Doesn't Work.

# Re-identification Demonstration Attack Summary

**What can we conclude from the empirical evidence provided by these 11 highly influential re-identification attacks?**

- The proportion of *demonstrated* re-identifications is extremely small.
- Which ***does not imply data re-identification risks are necessarily very small*** (*especially if the data has not been subject to Statistical Disclosure Limitation methods*).
- But with only 268 re-identifications made out of 327 million opportunities, Ohm's "Broken Promises" assertion that "*scientists have demonstrated they can often re-identify with astonishing ease*" seems rather dubious.
- It also seems clear that the state of "re-identification science", and the "evidence", it has provided needs to be dramatically improved in order to better support good public policy regarding data de-identification.

# *References for Re-identification Attack Summary Table*


1. Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
2. Barth-Jones, DC., The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now (July 2012). http://ssrn.com/abstract=2076397
3. Michael Barbaro, Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times August 6, 2006. www.nytimes.com/2006/08/09/technology/09aol.html
4. Narayanan, A., Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. Proceeding SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy p. 111-125.
5. Kwok, P.K.; Lafky,D. Harder Than You Think: A Case Study of Re-Identification Risk of HIPAA Compliant Records. Joint Statistical Meetings. Section on Government Statistics. Miami, FL Aug 2, 2011. p. 3826-3833.
6. El Emam K, et al. De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. J Med Internet Res 2012;14(1):e33
7. Valentino-DeVries, J. May the Best Algorithm Win… With $3 Million Prize, Health Insurer Raises Stakes on the Data-Crunching Circuit. Wall Street Journal. March 16, 2011. March 17, 2011 http://www.wsj.com/article_email/SB10001424052748704662604576202392747278936-lMyQjAxMTAxMDEwNTExNDUyWj.html
8. Narayanan, A. An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset. May 27, 2011 http://randomwalker.info/publications/heritage-health-re-identifiability.pdf
9. Narayanan, A. Felten, E.W. No silver bullet: De-identification still doesn't work. July 9, 2014 http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf
10. Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich. Identifying Personal Genomes by Surname Inference. Science 18 Jan 2013: 321-324.
11. Barth-Jones, D. Public Policy Considerations for Recent Re-Identification Demonstration Attacks on Genomic Data Sets: Part 1. Harvard Law, Petrie-Flom Center: Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations. http://blogs.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/
12. Sweeney, L., Abu, A, Winn, J. Identifying Participants in the Personal Genome Project by Name (April 29, 2013). http://ssrn.com/abstract=2257732

# *References for Re-identification Attack Summary Table*


13. Jane Yakowitz. Reporting Fail: The Reidentification of Personal Genome Project Participants May 1, 2013. https://blogs.harvard.edu/infolaw/2013/05/01/reporting-fail-the-reidentification-of-personal-genome-project-participants/
14. Barth-Jones, D. Press and Reporting Considerations for Recent Re-Identification Demonstration Attacks: Part 2. Harvard Law, Petrie-Flom Center: Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations. http://blogs.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/
15. Sweeney, L. Matching Known Patients to Health Records in Washington State Data (June 5, 2013). http://ssrn.com/abstract=2289850
16. Robertson, J. States' Hospital Data for Sale Puts Privacy in Jeopardy. Bloomberg News June 5, 2013. https://www.bloomberg.com/news/articles/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy
17. Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3, Article number: 1376 (2013) http://www.nature.com/articles/srep01376
18. Anthony Tockar. Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. September 15, 2014. https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/
19. Barth-Jones, D. The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore. https://blogs.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/
20. de Montjoye, et al. . Unique in the shopping mall: On the reidentifiability of credit card metadata. Science. 30 Jan 2015: Vol. 347, Issue 6221, pp. 536-539.
21. Barth-Jones D, El Emam K, Bambauer J, Cavoukian A, Malin B. Assessing data intrusion threats. Science. 2015 Apr 10; 348(6231):194-5.
22. de Montjoye, et al. Assessing data intrusion threats—Response Science. 10 Apr 2015: Vol. 348, Issue 6231, pp. 195
23. Jane Yakowitz Bambauer. Is De-Identification Dead Again? April 28, 2015. https://blogs.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/
24. David Sánchez, Sergio Martínez, Josep Domingo-Ferrer. Technical Comments: Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". Science. 18 Mar 2016: Vol. 351, Issue 6279, pp. 1274.
25. Sánchez, et al. Supplementary Materials for "How to Avoid Reidentification with Proper Anonymization"- Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". http://arxiv.org/abs/1511.05957
26. de Montjoye, et al. Response to Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata" Science 18 Mar 2016: Vol. 351, Issue 6279, pp. 1274

## *References for Re-identification Attack Summary Table*


27. Nate Anderson. “Anonymized” data really isn’t—and here’s why not. Sep 8, 2009 http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/
28. Sorrell v. IMS Health: Brief of Amici Curiae Electronic Privacy Information Center. March 1, 2011. https://epic.org/amicus/sorrell/EPIC_amicus_Sorrell_final.pdf
29. Ruth Williams. Anonymity Under Threat: Scientists uncover the identities of anonymous DNA donors using freely available web searches. The Scientist. January 17, 2013. http://www.the-scientist.com/?articles.view/articleNo/34006/title/Anonymity-Under-Threat/
30. Kevin Fogarty. DNA hack could make medical privacy impossible. CSO. March 11, 2013. http://www.csoonline.com/article/2133054/identity-access/dna-hack-could-make-medical-privacy-impossible.html
31. Adam Tanner. Harvard Professor Re-Identifies Anonymous Volunteers in DNA Study. Forbes. Apr 25, 2013. http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/
32. Adam Tanner. The Promise & Perils of Sharing DNA. Undark Magazine. September 13, 2016. http://undark.org/article/dna-ancestry-sharing-privacy-23andme/
33. Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903. September 29, 2015. http://techscience.org/a/2015092903
34. David Sirota. How Big Brother Watches You With Metadata. San Francisco Gate. October 9, 2014. http://www.sfgate.com/opinion/article/How-Big-Brother-watches-you-with-metadata-5812775.php
35. Natasha Singer. With a Few Bits of Data, Researchers Identify ‘Anonymous’ People. New York Times. Bits Blog. January 29, 2015. http://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/

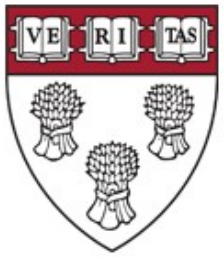

---

## *Additional Re-identification Attack Review References*


1. Khaled El Emam, Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. PLoS One 2011; Vol 6(12):e28071.
2. Jane Henriksen-Bulmer, Sheridan Jeary. Re-identification attacks - A systematic literature review. International Journal of Information Management, 36 (2016) 1184–1192.


Bill of Health

Examining the intersection of law and health care, biotech & bioethics
A blog by the Petrie-Flom Center and friends




# Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations

- http://blogs.law.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/

- https://blogs.law.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/

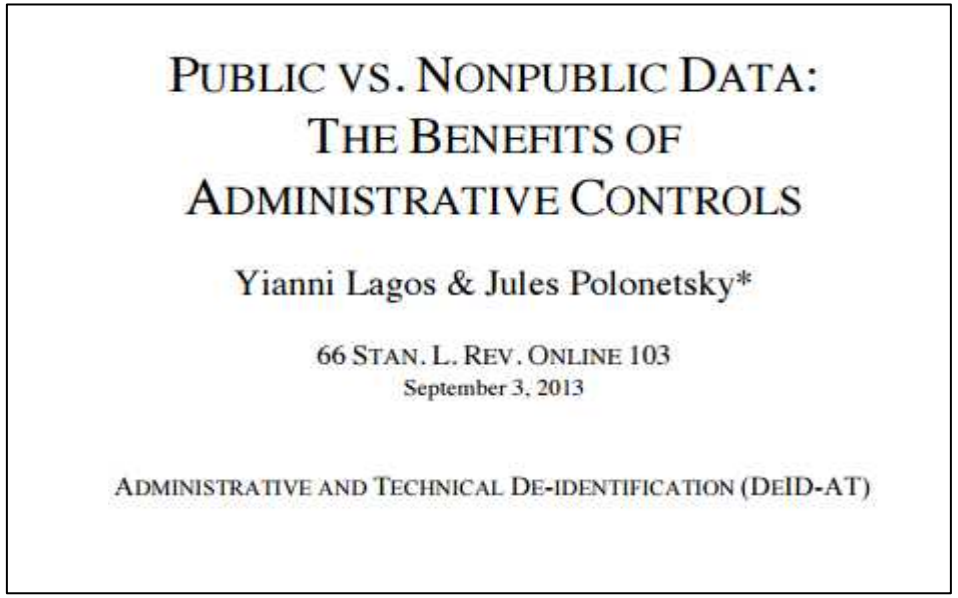
- http://blogs.law.harvard.edu/billofhealth/2013/10/02/ethical-concerns-conduct-and-public-policy-for-re-identification-and-de-identification-practice-part-3-re-identification-symposium/

# 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 Legislative Prohibitions Against Data Re-identification

## A BILL

To protect the privacy of potentially identifiable personal information by establishing accountability for the use and transfer of potentially identifiable personal information. [Version 4.4]

**SECTION 1. SHORT TITLE.**

This Act may be cited as the "Personal Data Deidentification Act".

**SEC. 2. DEFINITIONS.**

As used in this Act:

(1) DATA AGREEMENT.—The term "data agreement" means a contract, memorandum of understanding, data use agreement, or similar agreement between a discloser and a recipient relating to the use of personal information.

(2) DATA AGREEMENT SUBJECT TO THIS ACT.—The term "data

**Robert Gellman, 2010**
https://fpf.org/wp-content/uploads/2010/07/The_Deidentification_Dilemma.pdf

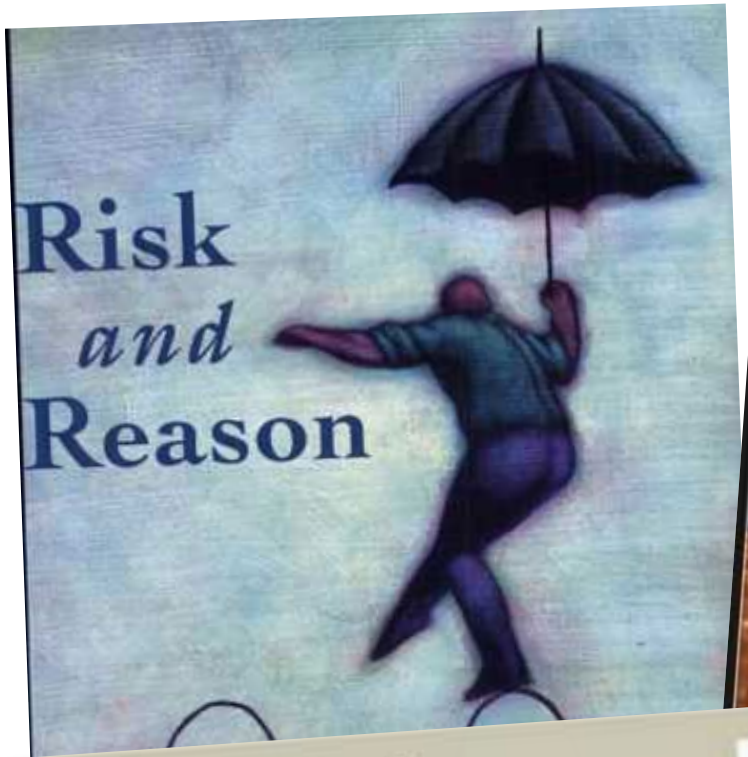
# HIPAA §164.514(b)(1)(i) and *Anticipated Recipients*

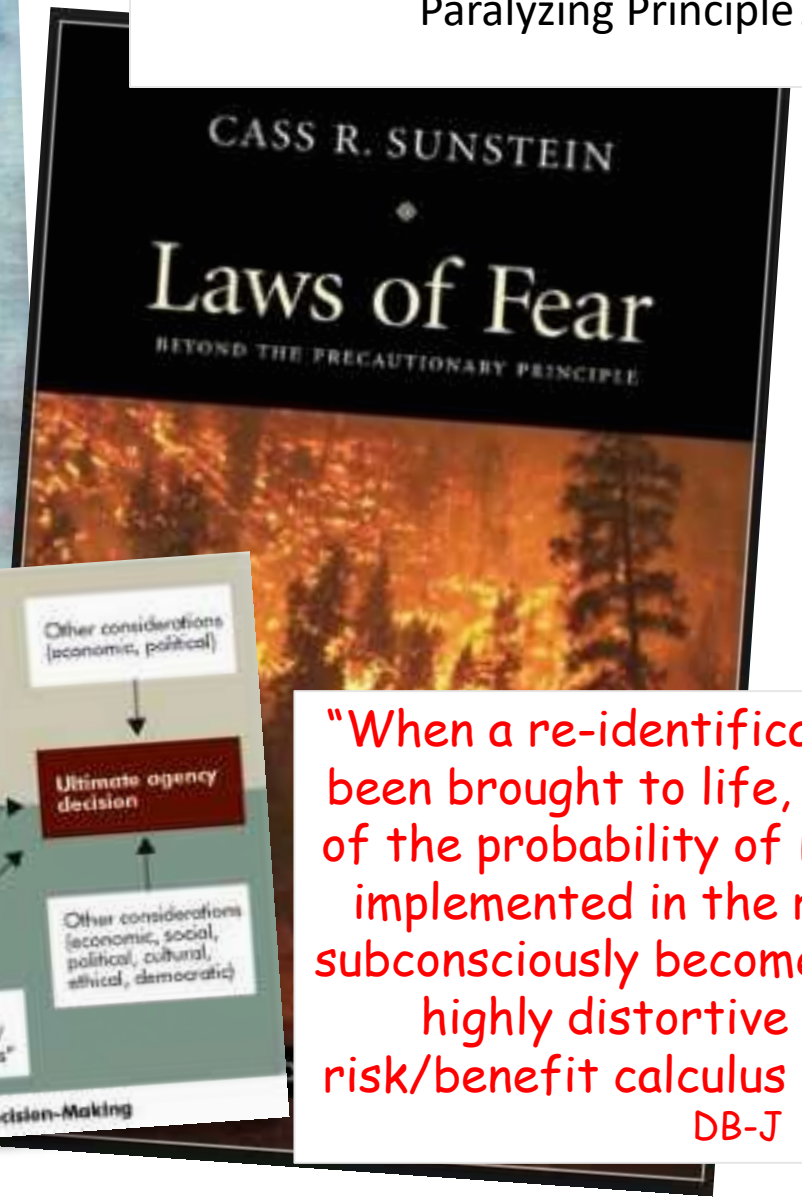
(i) Applying such principles and methods, determines that the *risk is very small* that *the information could be used*, alone or *in combination with other reasonably available information, by an anticipated recipient to identify an individual* who is a subject of the information;

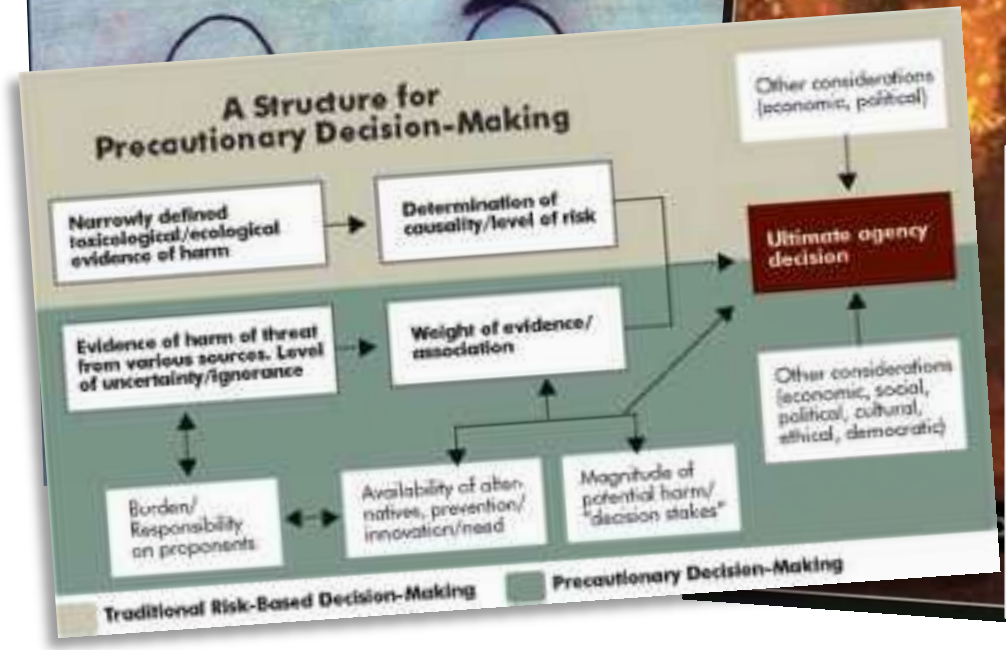
It is important to note that §164.514(b)(1)(i) is written with respect to "Anticipated Recipients". This introduces the concept of using policy, procedural and contract controls for limiting the Anticipated Recipients and the time periods and projects for which data is made available.

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

Precautionary Principle or
Paralyzing Principle?




"When a re-identification attack has been brought to life, our assessment of the probability of it actually being implemented in the real-world may subconsciously become 100%, which is highly distortive of the true risk/benefit calculus that we face." - DB-J

# 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).

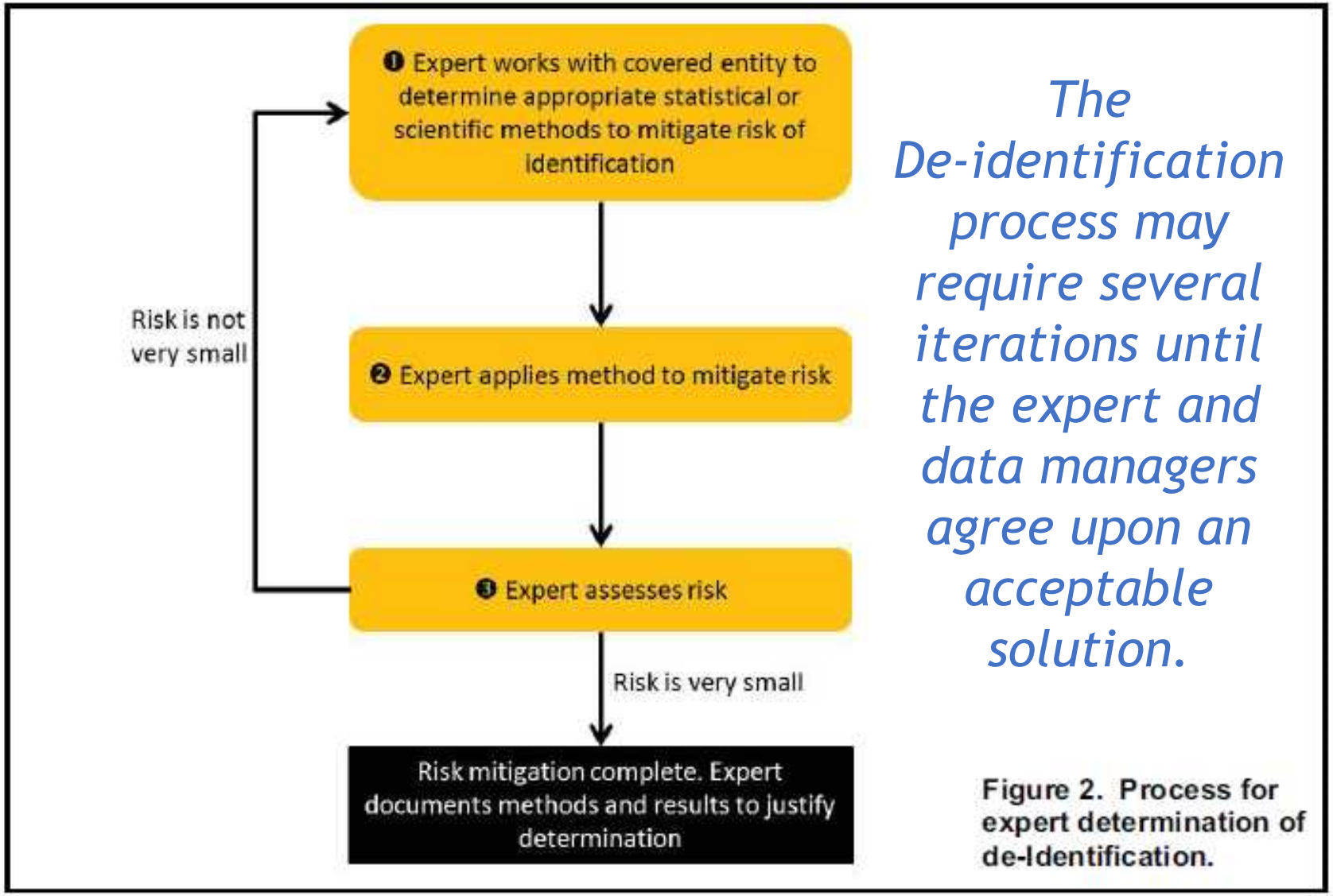
# Q2.5 *Can an expert derive multiple solutions from the same data set for a recipient?* *(Cont'd)*

- 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.

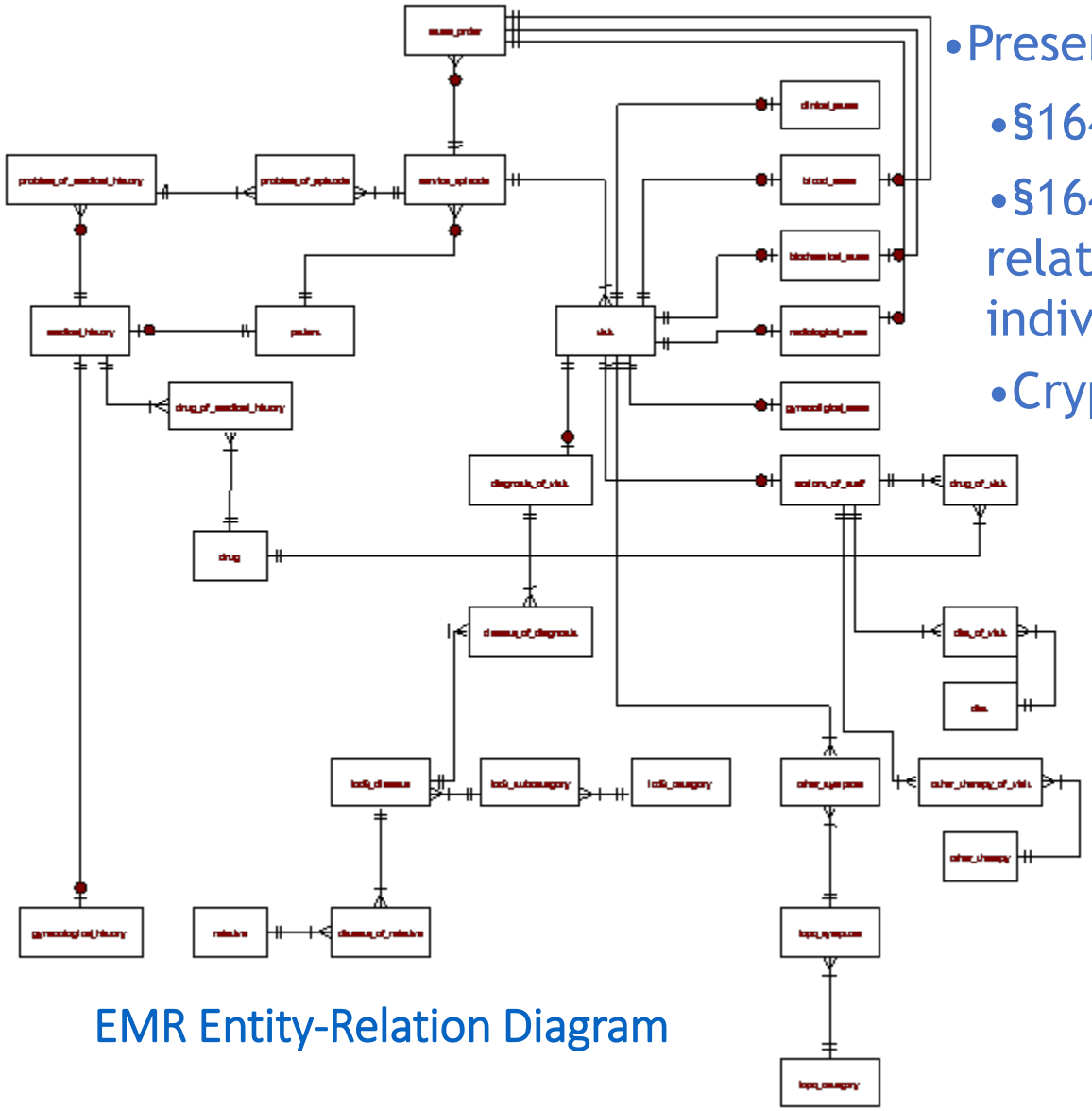
# Q2.8. *What are the approaches by which an expert mitigates the risk of identification?* *(Cont'd)*

- 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

# *Why Privacy Science Must Become A "Systems Science"*

- Paul Ohm described a dystopic vision that all information is effectively PII and that the failure of perfect de-identification would lead us through cycles of accretive re-identification toward a universal "database of ruin".
- This misconception ignores the underlying mathematical realities which indicate that when modern statistical disclosure limitation (SDL) methods can be used to effectively de-identify data, we will have resulting increases in "false positive" re-identifications.
- Such false positive linkages will practically prevent the ability of such systemic "crystallization" of iteratively linked de-identified data into accurate dossiers for the very vast majority of the population.
- Because of this de-identification, although imperfectly protective, is critical for reaching reasonable solutions which can continue to offer pragmatic and sustainable data obscurity in the evolving era of big data.

*Brussels Privacy Symposium:*
*Identifiability: Policy and Practical Solutions for Anonymization and Pseudonymization.*

# Why a Systems-Science Perspective is Needed to Better Inform Data Privacy De-identification Public Policy, Regulation and Law

Daniel C. Barth-Jones, M.P.H., Ph.D.
Department of Epidemiology
Mailman School of Public Health
Columbia, University

## Introduction

Systems sciences pursue an understanding of how parts (including individual actors and groups) within a larger system combine via their interactions to produce emergent phenomena which are greater than, or different from, a mere sum of the parts. Important examples of systems behaviors include the existence of threshold phenomena, which allows infectious diseases to spread as epidemics through a population, or the sudden crystallization of supersaturated solutions. I argue that data privacy policy for de-identification must take a systems perspective in order to better understand how combined multi-dimensional (i.e., involving both technical de-identification and administrative/regulatory responses) interventions can effectively combine to create practical controls for countering wide-spread re-identification threats.

# *Why Privacy Science Must Become A "Systems Science"*

- Modern SDL-based de-identification essential protections for preventing mass re-identification at scale and positions advocating for wholesale abandonment of de-identification due to less-than-perfect efficacy discard one of data privacy's most effective tools for an idealistic hope of perfect privacy protections makes "perfect the enemy of the good".
- Systems perspective using uncertainty analyses can help to apply consistent and rigorous probabilistic methods accounting for our uncertainty about the efficacy of various technical, administrative and legal protections at different stages in data intrusion scenarios to demonstrate that combining these methods can lead to useful assurance that (admittedly less than perfect) de-identification can still provide useful protections without resorting to only worst case scenarios about data intruder's knowledge.

# *Re-identification Science Policy Short-comings:*

6 ways in which "Re-identification Science" has (thus far) typically failed to best support sound public policies:

1) Attacking only trivially "straw man" de-identified data, where modern statistical disclosure control methods (like HIPAA) weren't used.

2) Targeting only especially vulnerable subpopulations and failing to use statistical random samples to provide policy-makers with representative re-identification risks for the entire population.

3) Making bad (often worst-case) assumptions and then failing to provide evidence to justify assumptions.

Corollary: Not designing experiments to show the boundaries where de-identification finally succeeds.

# *Re-identification Science Policy Short-comings:*

6 ways in which "Re-identification Science" has (thus far) typically failed to support sound public policies (Cont'd):

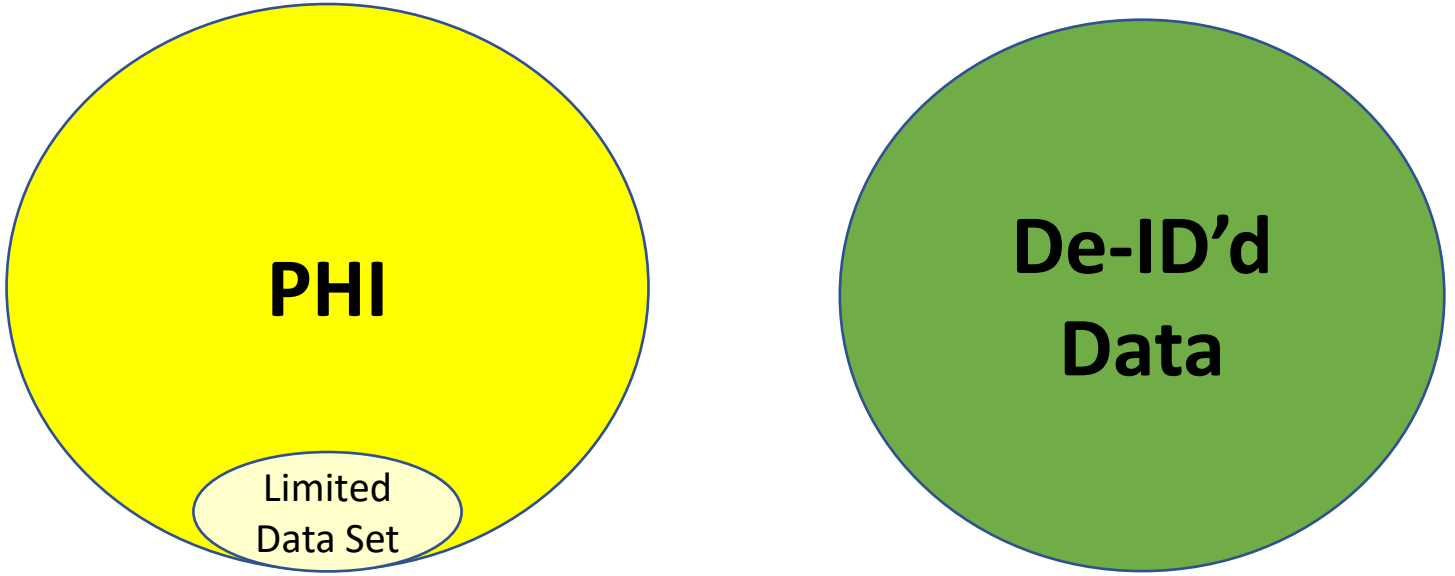
4) Failing to distinguish between sample uniqueness, population uniqueness and re-identifiability (i.e., the ability to correctly link population unique observations to identities).

5) Failing to fully specify relevant threat models (using data intrusion scenarios that account for all of the motivations, process steps, and information required to successfully complete the re-identification attack for the members of the population).

6) Unrealistic emphasis on absolute "Privacy Guarantees" and *failure to recognize unavoidable trade-offs between data privacy and statistical accuracy/utility*.

# De-identification under HIPAA - Basics

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




*De-ID'd data is outside HIPAA*
*Contract requirements may apply*