Responsible Sharing of Health Information

Khaled El Emam
kelemam@ehealthinformation.ca

11th October 2022

Disclosures

- Khaled El Emam is co-founder and director of Replica Analytics Ltd, a spinoff company from the University of Ottawa / CHEO Research Institute specializing in the development of data synthesis software for health data. In December 2021 Replica was acquired by Aetion.

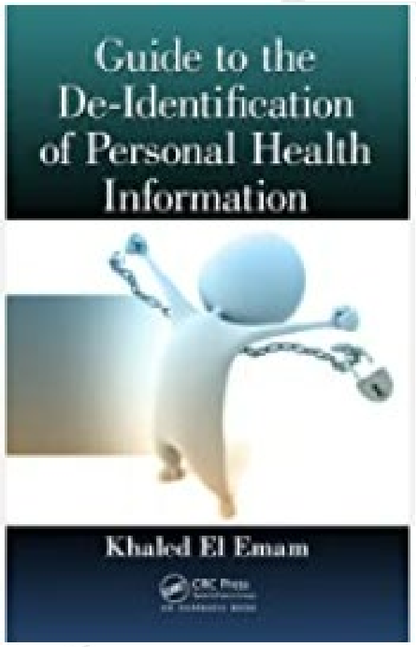
Agenda for Today

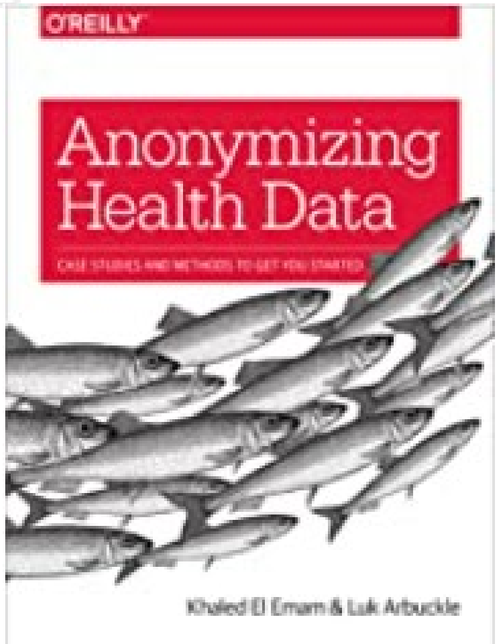
Basic concepts 1 Basic concepts to consider when sharing health data

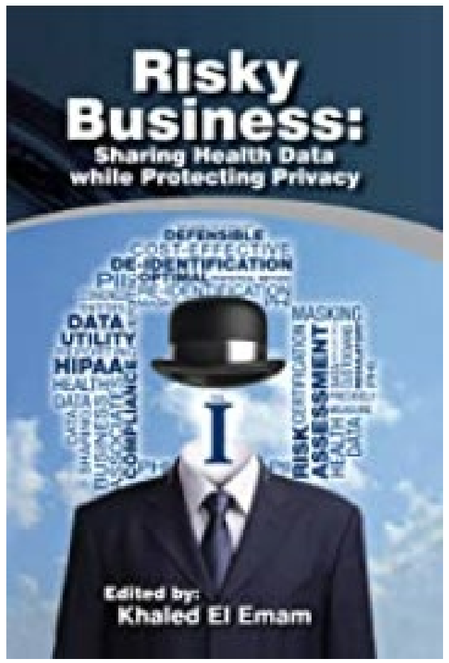
Re-identification risk measurement 2 An overview of methods for risk measurement

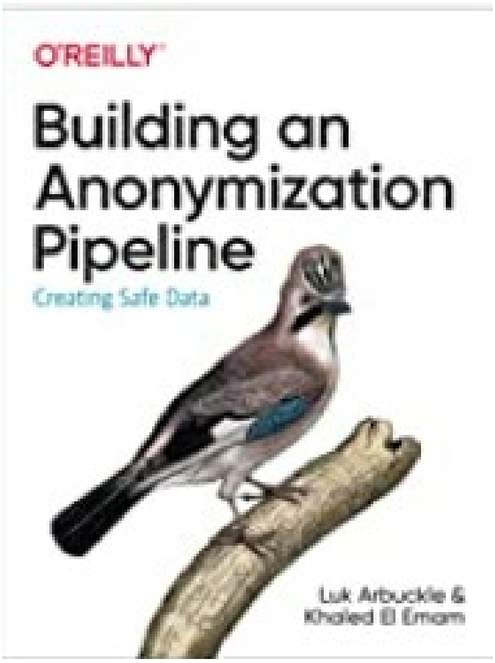
The future 3 Where things are headed with data sharing

We have authored or co-authored a series of books on the topic









Obligations on processing personal information

- There are many obligations on the processing of personal health information (PHI), including the requirement to obtain data subject consent / authorization
- The consent often needs to be specific to a particular purpose and the PHI cannot be used for a different purpose unless further consent is obtained
- Use of personal data for the consented purpose is deemed to be a **primary purpose**
- The scope of how personal data can be used based on a specific consent can vary by jurisdiction, for example, an organization can make a legitimate interest argument
- There are some exceptions to use and disclosure of PHI without consent, such as for reporting communicable disease, for example

Obligations on processing personal information

- Otherwise, the use and disclosure of personal data would be for a **secondary purpose**, for which consent was not obtained and there is no exception
- In general, if data is rendered to be non-identifiable (i.e., de-identified) then no consent is required
- De-identified information is not considered to be PHI; it is not considered to be personal information and therefore can often fall outside privacy statutes, or can be processed with fewer obligations

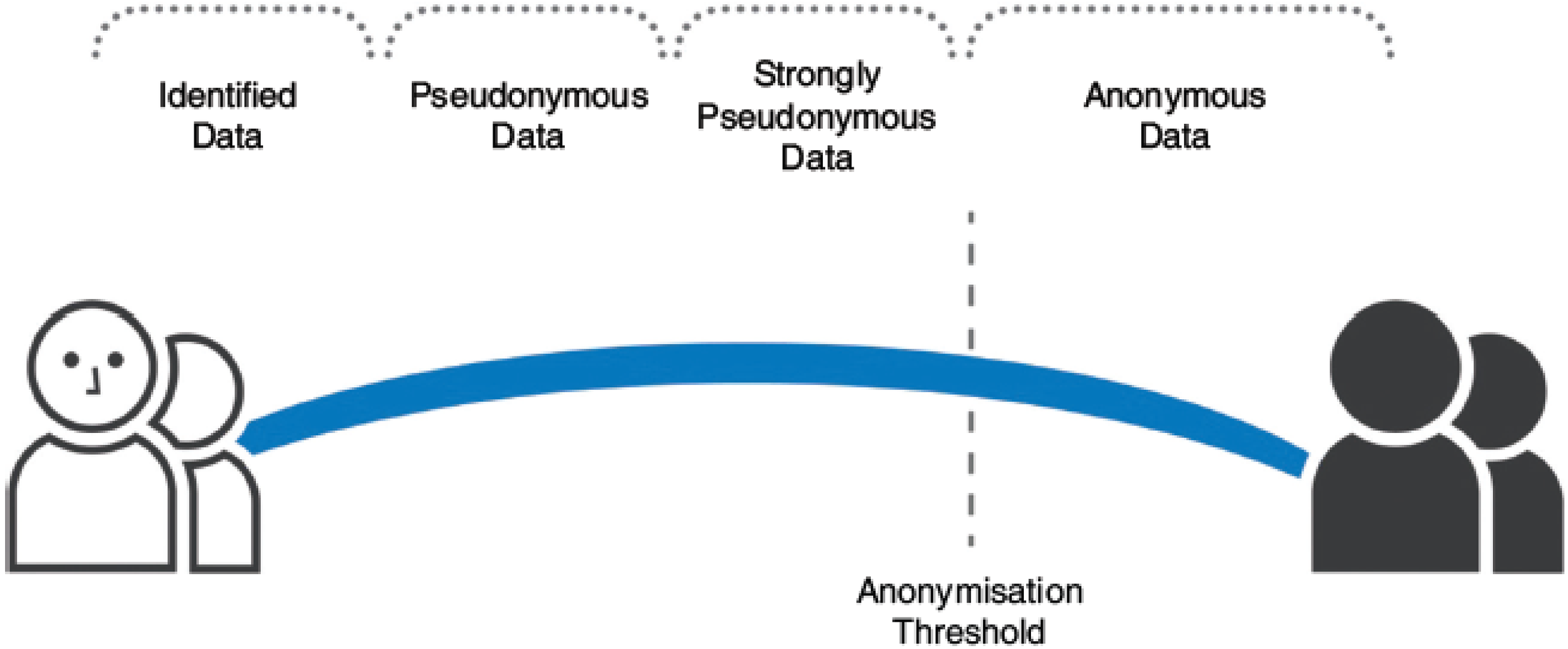
Primary purposes

- Purposes related to the provision of care are considered to be primary purposes; this includes using and disclosing data by/to the individuals involved in the circle of care
- Other purposes such as billing and processing insurance payments are also often considered to be primary purposes
- What is a primary purpose is a legal question, however, and if there is ambiguity then reference to relevant legislation is advised / legal advice should be sought

Secondary purposes

- Secondary purposes include research and public health
- Also, obvious data uses, such as building models for marketing purposes are secondary purposes
- In general, testing software applications are increasingly being seen as secondary purposes as well
- Open data, unless explicitly stated in the consent when the data was collected, would be considered a form of secondary processing as well

The different states of data



Each data state has certain obligations

GDPR obligation	Type of data			
	Identified	**Pseudonymised (basic)**	**Strongly pseudonymised**	**Anonymised**
1. Provide notice to data subject	Required	Required	Required	Not required
2. Legal basis for processing (legitimate interests, consent)	Required	Stronger case for legitimate interests	Much stronger case	Not required
3. Data subject rights (access, portability, rectification)	Required	Required	Not required	Not required
4. Give right to erasure/right to be forgotten	Required	Required	May not be required	Not required
5. Basis for cross-border transfers	Required	Required	Required	Not required
6. Data protection by design	Required	Partially met	Strengthens the ability to meet this obligation	Not required
7. Data security	Required	Partially met	Strengthens the ability to meet this obligation	Not required
8. Data breach notification	Likely to be required	Less likely to be required	Strengthens the case that notification is not required	Not required
9. Data retention limitations	Required	Required	Required	Not required
10. Documentation/recordkeeping obligations	Required	Required	Required	Not required
11. Vendor/sub-processor management	Required	Required	Required	Not required

M. Hintze and K. El Emam, “Comparing the benefits of pseudonymisation and anonymisation under the GDPR,” J. Data Prot. Priv., vol. 2, no. 1, pp. 145–158, Dec. 2018.

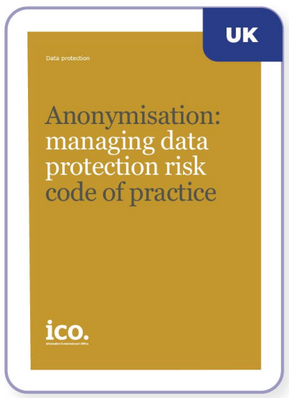
De-identification of PHI

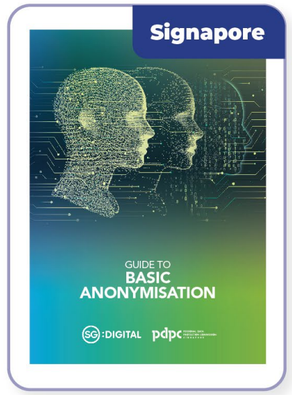
- The use and disclosure of data for secondary purposes can be enabled by de-identification
- This includes data transfers across jurisdictions
- One of the main objectives of de-identification is to protect against identity disclosure
- It is a risk management exercise in that it is intended to ensure that the risk of identity disclosure is very small
- In general, the act of de-identification does not require additional consent; the reasoning will depend on the statute that is applicable

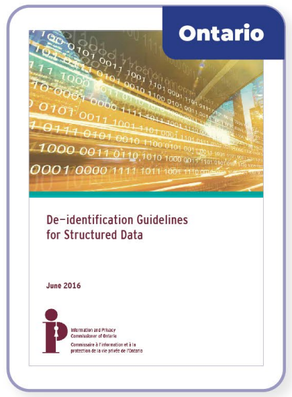
Obligations on processing de-identified PHI

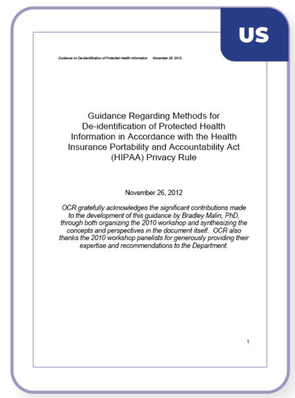
- While the obligations on de-identified PHI are reduced, they are not completely zero
- There is increasingly a prohibition against re-identification
- Risk is managed by data transformations and additional controls – it is necessary to ensure that the controls travel with the data

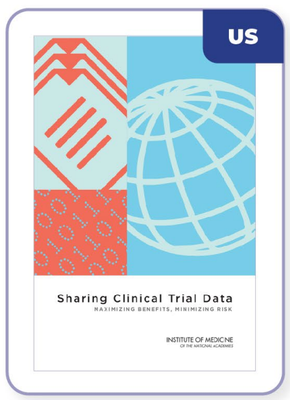
There is a consistent approach in existing standards and guidelines

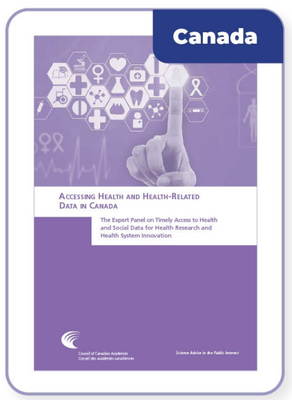


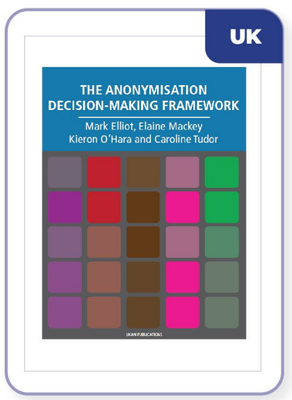












DIRECT IDENTIFIERS

- Name
- Email address
- SIN / SSN
- Biometrics
- Health insurance number
- Full residential address

INDIRECT IDENTIFIERS

- Postal code / ZIP code
- Age / DoB
- Race / ethnicity / language
- Income
- Visible characteristics (e.g., mobility devices)
- Dates of important events (e.g., marriage, death

Basic definitions – identity disclosure is when a person's identity is assigned to a record

Quasi-identifiers

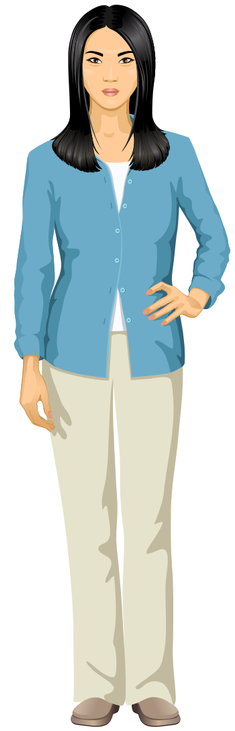
Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

Basic definitions – generalization means that more than one record can match a person

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	1980-1989	65862-403
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Attacks can be in two directions – population to sample attack

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	**1980-1989**	**55566-2110**
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540



Attacks can be in two directions – sample to population attack

Sex	Year of Birth	NDC
Male	1970-1979	009-0031
Male	1980-1989	0023-3670
Male	1970-1979	0074-5182
Female	1990-1999	0078-0379
Female	**1980-1989**	**65862-403**
Male	1990-1999	55714-4446
Male	1990-1999	55714-4402
Female	1980-1989	55566-2110
Male	1970-1979	55289-324
Female	1990-1999	54868-6348
Male	1980-1989	53808-0540

Risk is measured by the group size

Sex	Year of Birth	NDC	Group Size	Risk
Male	1975	009-0031	1	1
Male	1988	0023-3670	1	1
Male	1972	0074-5182	1	1
Female	1993	0078-0379	1	1
Female	**1989**	**65862-403**	**1**	**1**
Male	1991	55714-4446	1	1
Male	1992	55714-4402	1	1
Female	1987	55566-2110	1	1
Male	1971	55289-324	1	1
Female	1996	54868-6348	1	1
Male	1980	53808-0540	1	1

When we generalize the group size gets bigger, so the risk decreases

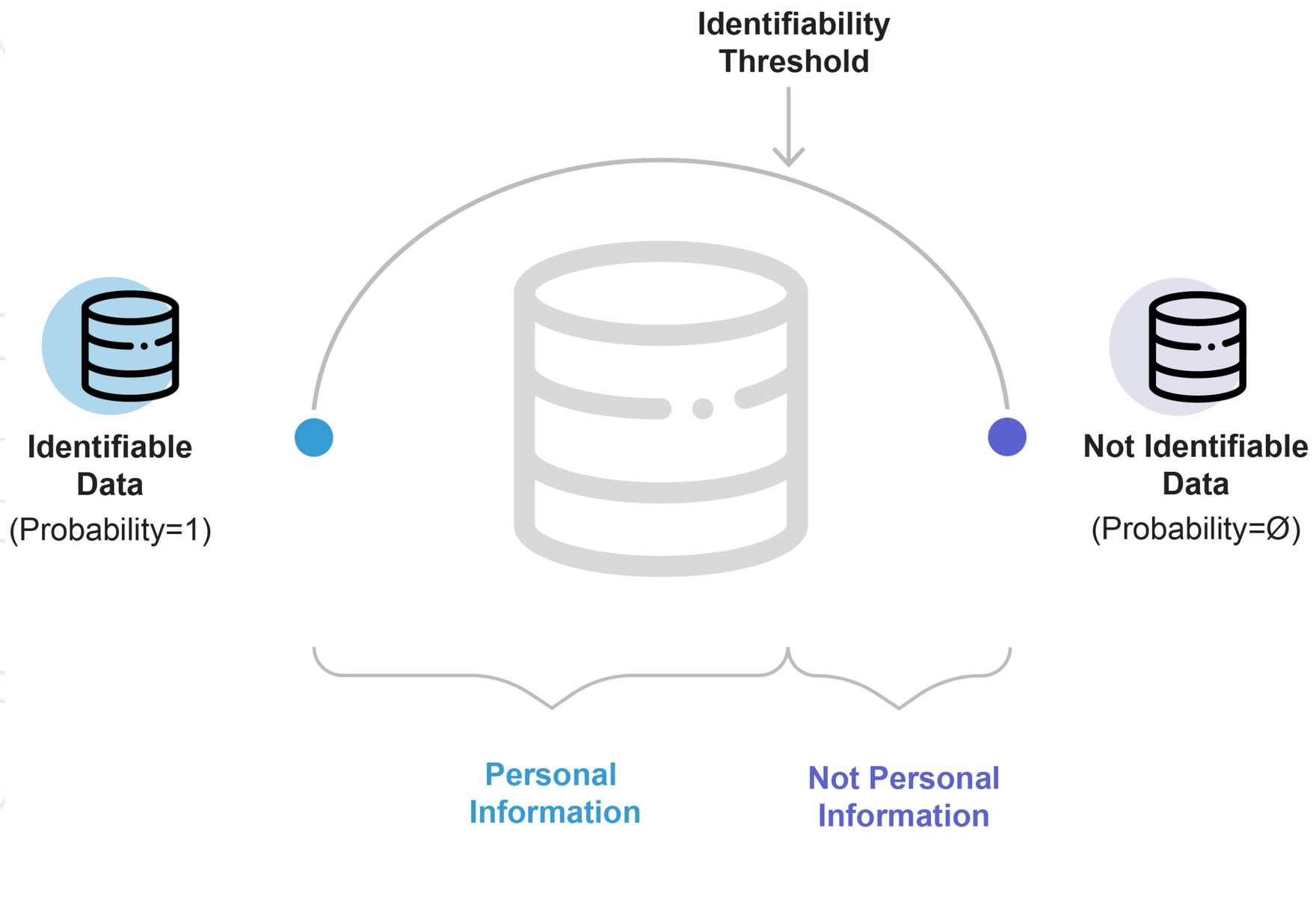
Sex	Decade of Birth	NDC	Group Size	Risk
Male	1970-1979	009-0031	3	0.33
Male	1980-1989	0023-3670	2	0.5
Male	1970-1979	0074-5182	3	0.33
Female	1990-1999	0078-0379	2	0.5
Female	**1980-1989**	**65862-403**	**2**	**0.5**
Male	1990-1999	55714-4446	2	0.5
Male	1990-1999	55714-4402	2	0.5
Female	**1980-1989**	**55566-2110**	**2**	**0.5**
Male	1970-1979	55289-324	3	0.33
Female	1990-1999	54868-6348	2	0.5
Male	1980-1989	53808-0540	2	0.5

But it is actually the population group size that matters

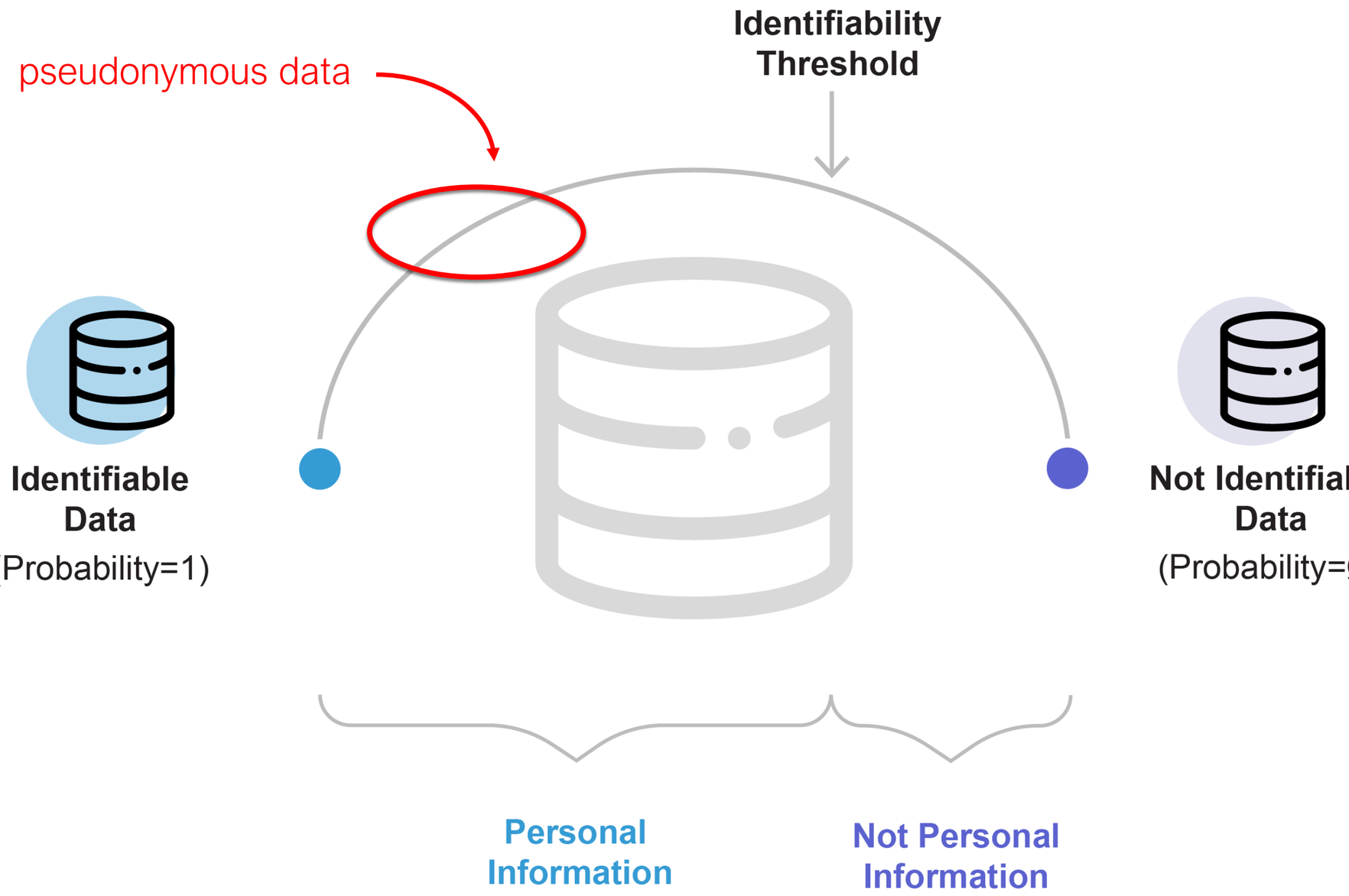
Sex	Decade of Birth	NDC	Group Size	Risk
Male	1970-1979	009-0031	3	
Male	1980-1989	0023-3670	2	
Male	1970-1979	0074-5182	3	
Female	1990-1999	0078-0379	2	
Female	**1980-1989**	**65862-403**	**2**	**0.1**
Male	1990-1999	55714-4446	2	
Male	1990-1999	55714-4402	2	
Female	**1980-1989**	**55566-2110**	**2**	**0.1**
Male	1970-1979	55289-324	3	
Female	1990-1999	54868-6348	2	
Male	1980-1989	53808-0540	2	

N=10

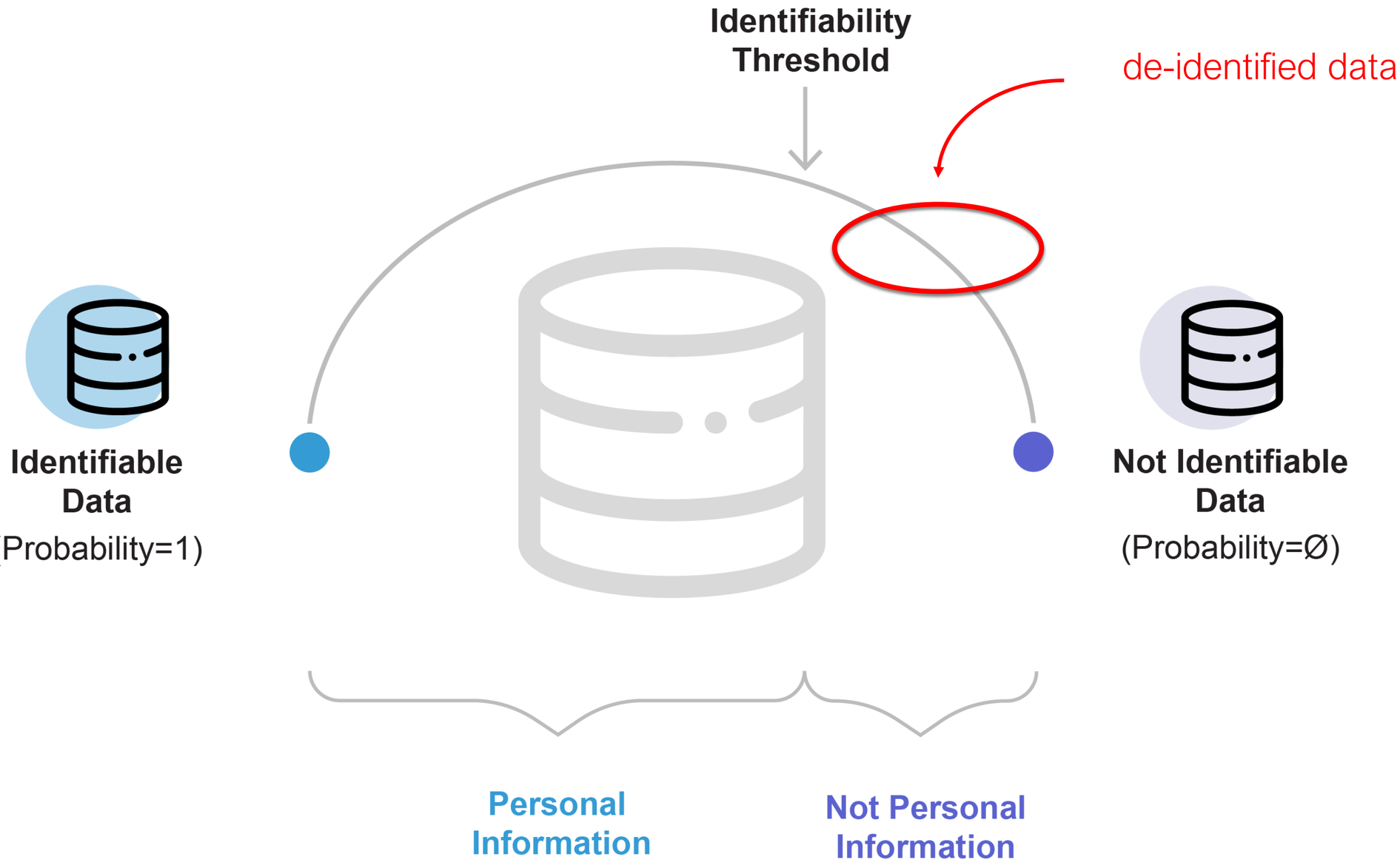
Identifiability spectrum and risk thresholds



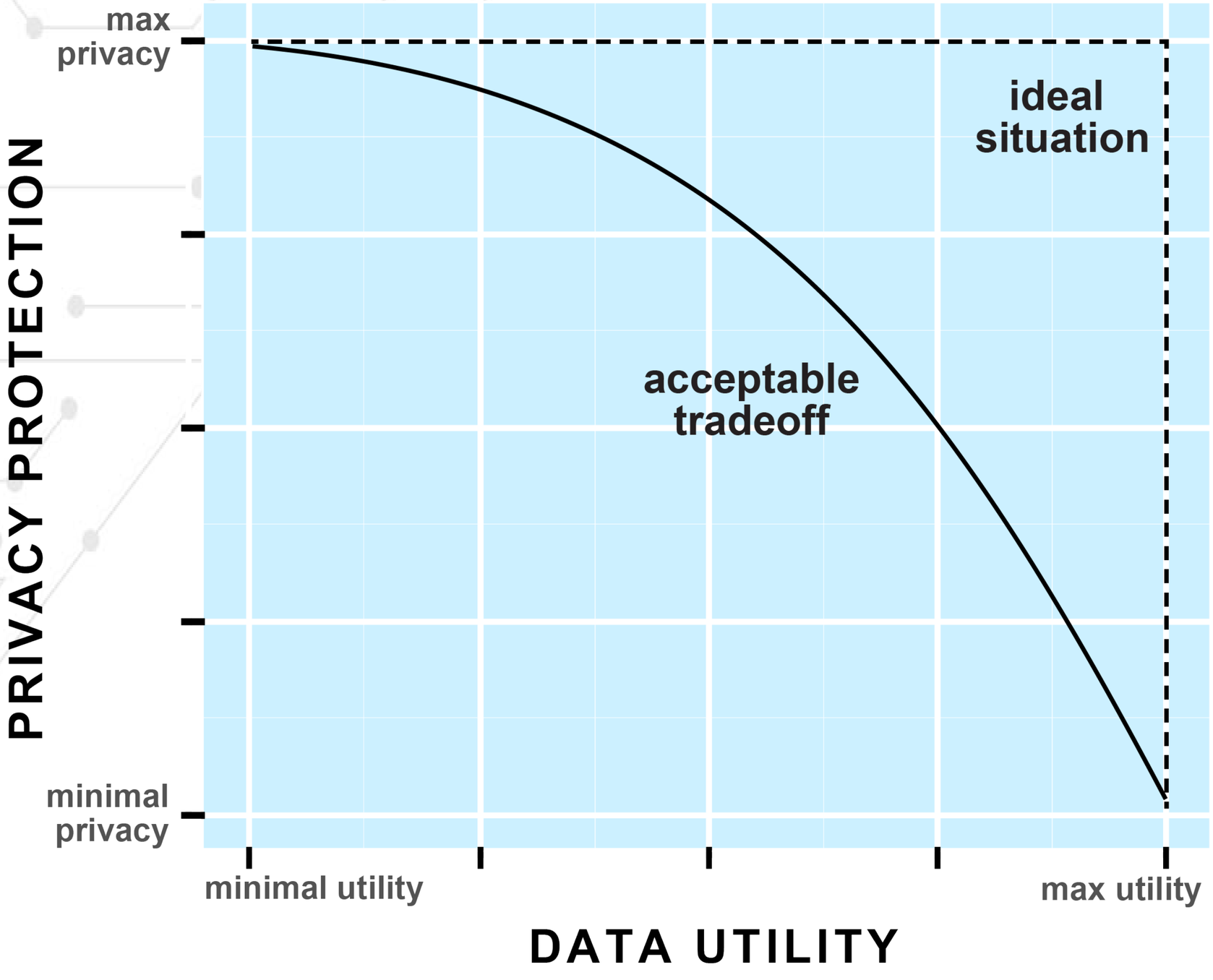
Pseudonymous data on the spectrum



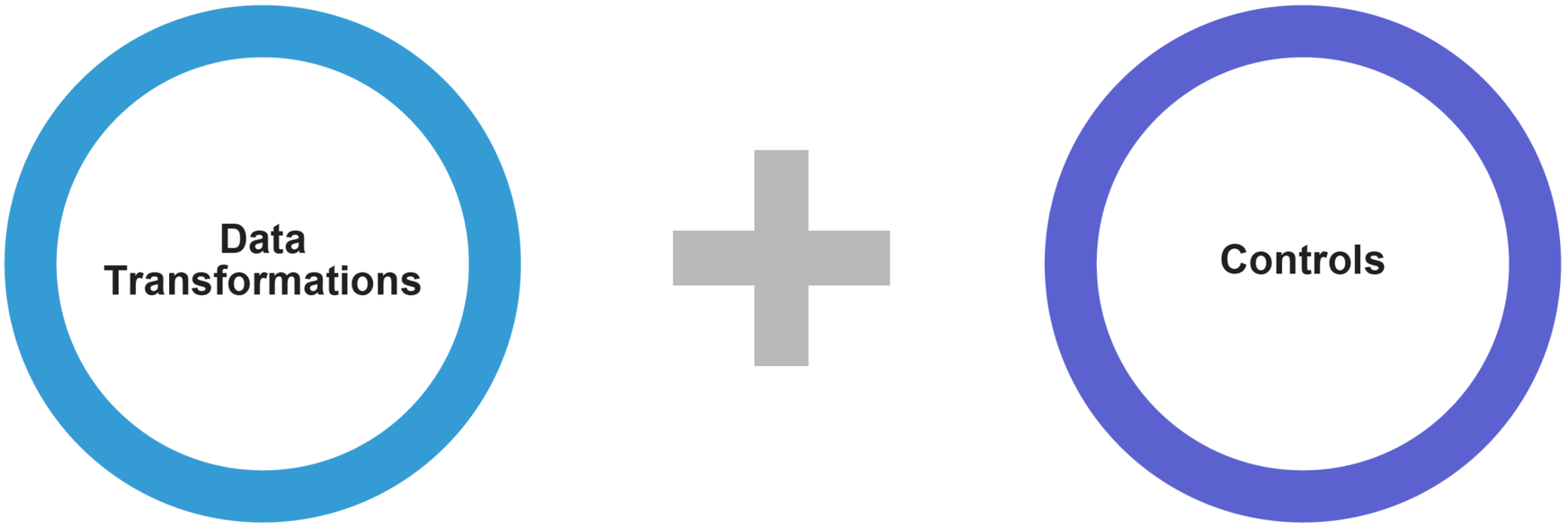
De-identified data on the spectrum



Privacy-Utility Trade-off



A common approach that has worked well in practice is risk-based anonymization



- Generalization
- Suppression
- Addition of noise
- Microaggregation

- Security controls
- Privacy controls
- Contractual controls

Claims of successful re-identification attacks, while debatable, still have created a negative narrative around traditional anonymization methods

The New York Times

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

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

Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel

DEC. 19, 2019

theguardian

'Anonymised' data can never be totally anonymous, says study

Findings say it is impossible for researchers to fully protect real identities in datasets

You're very easy to track down, even when your data has been anonymized

A new study shows you can be easily re-identified from almost any database, even when your personal details have been stripped out.

by **Charlotte Jee** Jul 23, 2019

ACM TECHNEWS

'Anonymized' Data Can Never Be Totally Anonymous, says Study

By The Guardian

HUFFPOST

Online Profiling and Invasion of Privacy: The Myth of Anonymization

02/20/2013 12:23 pm ET | **Updated** Apr 22, 2013

Electronic Health Information Laboratory

Commonly mentioned privacy enhancing technologies

01

RISK-BASED DE-IDENTIFICATION

Using methods like k-anonymity to measure re-identification risk, and data transformations are combined with controls to manage overall risk.

02

SYNTHETIC DATA GENERATION

Models are built from data, and these are used to generate new datasets that retain the statistical patterns.

03

FEDERATED ANALYSIS/SECURE MULTIPARTY COMPUTATION

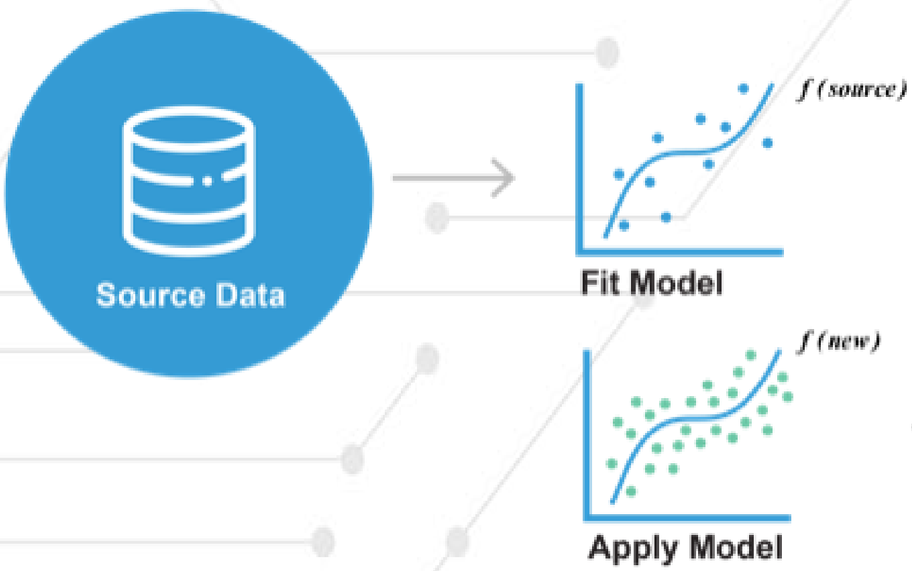
Computations are distributed among multiple parties, either as data sources or as computing nodes, or both.

04

DIFFERENTIAL PRIVACY

Interactive system that adds noise to the results of interactive queries to manage re-identification risk.

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

Additional Clarifications

- The source datasets can be as small as 100 or 150 patients. We have developed generative modeling techniques that will work for small datasets.
- The source datasets can be very large – then it becomes a function of compute capacity that is available.
- It is not necessary to know how the synthetic data will be analyzed to build the generative models. The generative models capture many of the patterns in the source data.


?
QUESTIONS