Practical De-identification Methods

VIPSS 2026

Electronic Health Information Laboratory, Children's Hospital of Eastern Ontario Research Institute

Workshop Agenda

1.30 – 1.45	Introduction
1.45 – 2.15	Scope and Terminology in this workshop
2.15 – 2.45	Concepts in Risk-Based De-Identification
2.45 – 3.00	Demo 1: Automated Risk Assessment
3.00 – 3.15	Break
3.15 – 3.45	Measuring Vulnerability
3.45 – 4.30	Modeling the Attacks
4.30 – 4.45	Demo 2: Understanding The Risk Report

Practical De-Identification Methods

SCOPE AND TERMINOLOGY

Module Agenda

Overview of the workshop — 1 — Learning objectives and scope of this workshop

Terminology — 2 — Definitions of important terms that we will use in this workshop

Setting thresholds — 3 — Thresholds for identity disclosure risk

uOttawa

Scope and Terminology

OVERVIEW OF THE WORKSHOP

Learning Objectives

- Get a broad understanding of the considerations around managing disclosure risks in data
- Learn practical methods for evaluating disclosure risks, and for managing them
- Be able to judge what are good practices for de-identification and for managing disclosure risks in general
- Be prepared to explain and justify practices for evaluating and managing disclosure risks

This workshop is not intended to provide legal advice.

uOttawa

Why De-Identify ?

Enable Secondary Use

Permit uses and disclosures of data for research, public health, and policy purposes beyond the original collection context.

Data Minimization

Reduce the amount of identifiable information retained or shared, limiting exposure consistent with privacy-by-design principles.

Data Disposal

Serve as a form of data deletion when full removal is not feasible, effectively rendering residual data non-identifiable.

AI Model Sharing

Enable the responsible sharing of AI/ML models trained on personal data — a growing challenge as regulators scrutinize model anonymity claims.

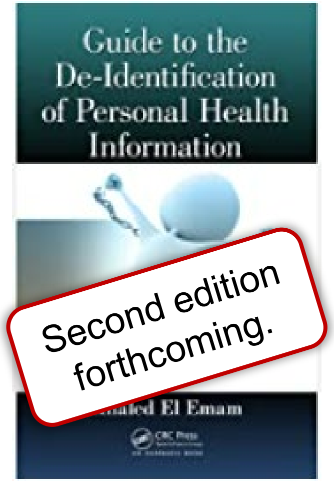
Scope of Methodology and Workshop

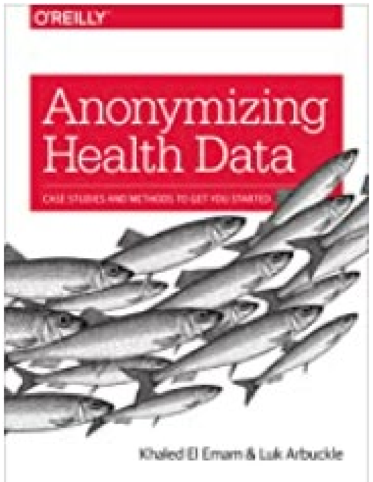
- **Data Context:**
  - Structured, individual-level tabular data
  - Excludes unstructured text, images and audio (although core principles apply)
  - Excludes aggregate data such as tables of counts
  - Assumes data already collected
- **Analytical Approach:**
  - Quantitative model of risk
  - Focus on measurable risk metrics and defensible thresholds
  - Focus on the risk side of the risk-utility framework
  - Conservative where multiple options exist

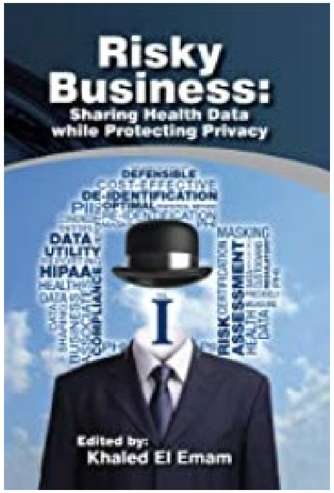
uOttawa

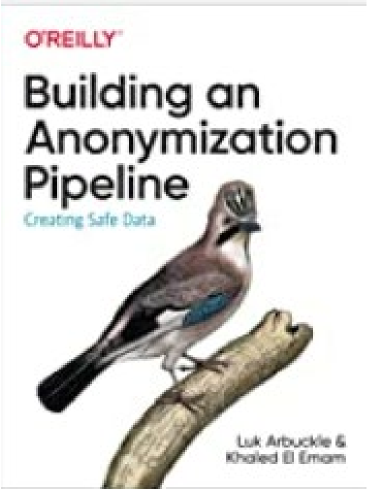
Scientific and Empirical Foundations

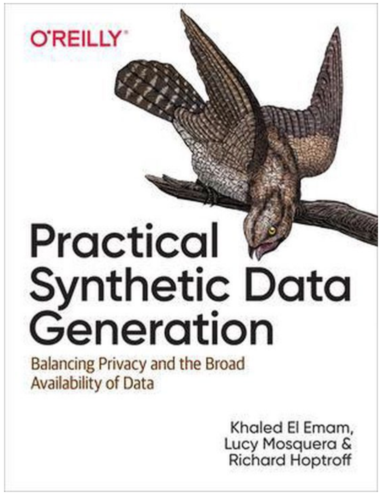
- Grounded in documented re-identification attacks and regulatory enforcement actions
- Based on peer-reviewed body on research (15+ years) and continuously evolving with new technologies and new attacks





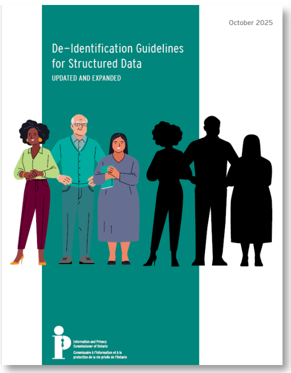




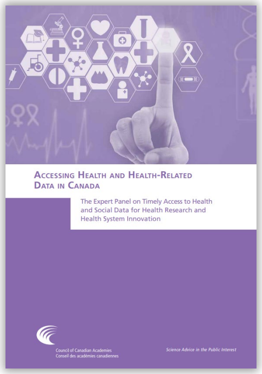


uOttawa

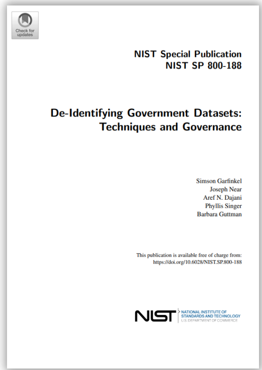
Standards and Frameworks Informing This Approach



Canada

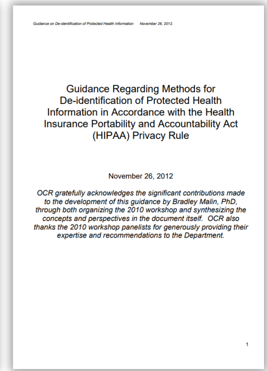


Canada

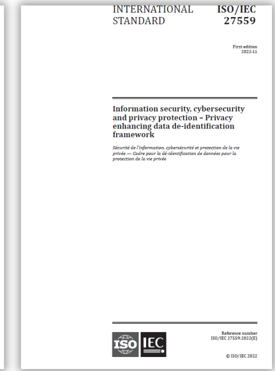


USA

USA

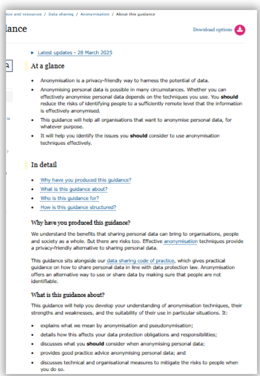


USA

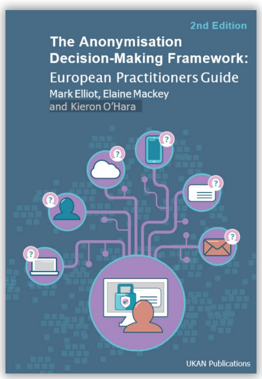


International

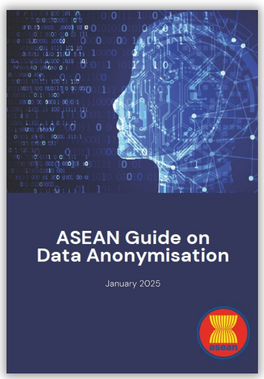
Israel



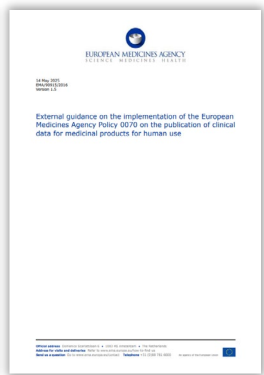
UK



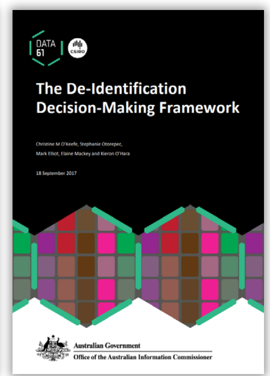
UK



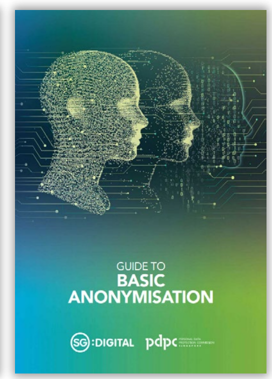
ASEAN



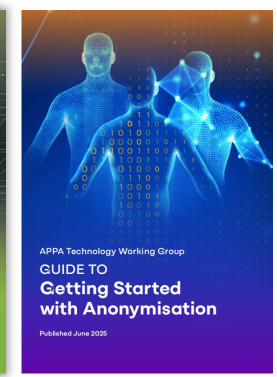
Europe



Australia



Singapore



APPA

Scope and Terminology

TERMINOLOGY

Terminology for This Workshop

- **Pseudonymization:** process of transforming direct identifiers that exist within a dataset[1]. The noun is pseudonymized data or pseudonymous data.

- **De-identification:** process of performing pseudonymization, plus transforming indirect identifiers that remain in the dataset following pseudonymization[1]. It also involves appropriate controls to reduce the overall risk.

[1] From the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

Terminology Varies Across Jurisdictions and Parties

- Inconsistent terminology across domains and jurisdictions
- Identifiability is defined and operationalized differently
- The terms *de-identification*, *pseudonymization* and *anonymization* are used inconsistently

Reducing identifiability in cross-national perspective: Statutory and policy definitions for anonymization, pseudonymization, and de-identification in G7 jurisdictions

11 October 2024

Interest in the use and application of processes and technologies for reducing the identifiability of individuals from their personal information [1] has accelerated in recent years. This includes technologies for de-identifying, pseudonymizing, and anonymizing personal information.

When applied appropriately, these processes and technologies can facilitate innovative uses of data, help

From: https://www.priv.gc.ca/en/opc-news/news-and-announcements/2024/de-id_20241011/

uOttawa

The Term *Anonymization* Under the European GDPR

Criterion	Definition
Singling Out	which corresponds to the possibility to isolate some or all records which identify an individual in the dataset;
Linkability	which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g. by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against "singling out" but not against linkability;
Inference	which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

Risks Essential to Anonymization According to the Opinion 05/2014 by the Article 29 Protection Working Party

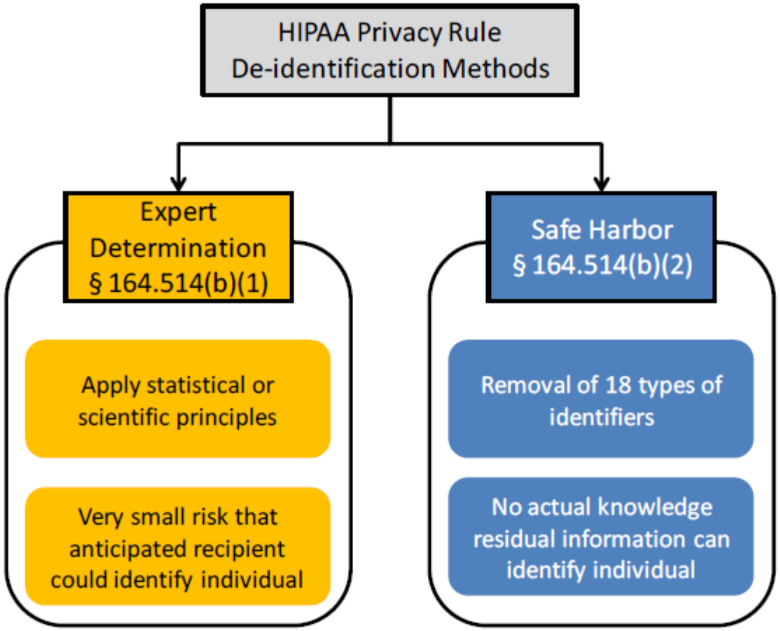
The Term *Anonymization* Under the Quebec Regulation

Criterion	Definition
Correlation	means the inability to connect datasets concerning the same person;
Individualization	means the inability to isolate or distinguish a person within a dataset;
Inference	means the inability to infer personal information from other available information.

Criteria Applicable to Anonymization According to the Regulation Respecting the Anonymization of Personal Information in Quebec

uOttawa

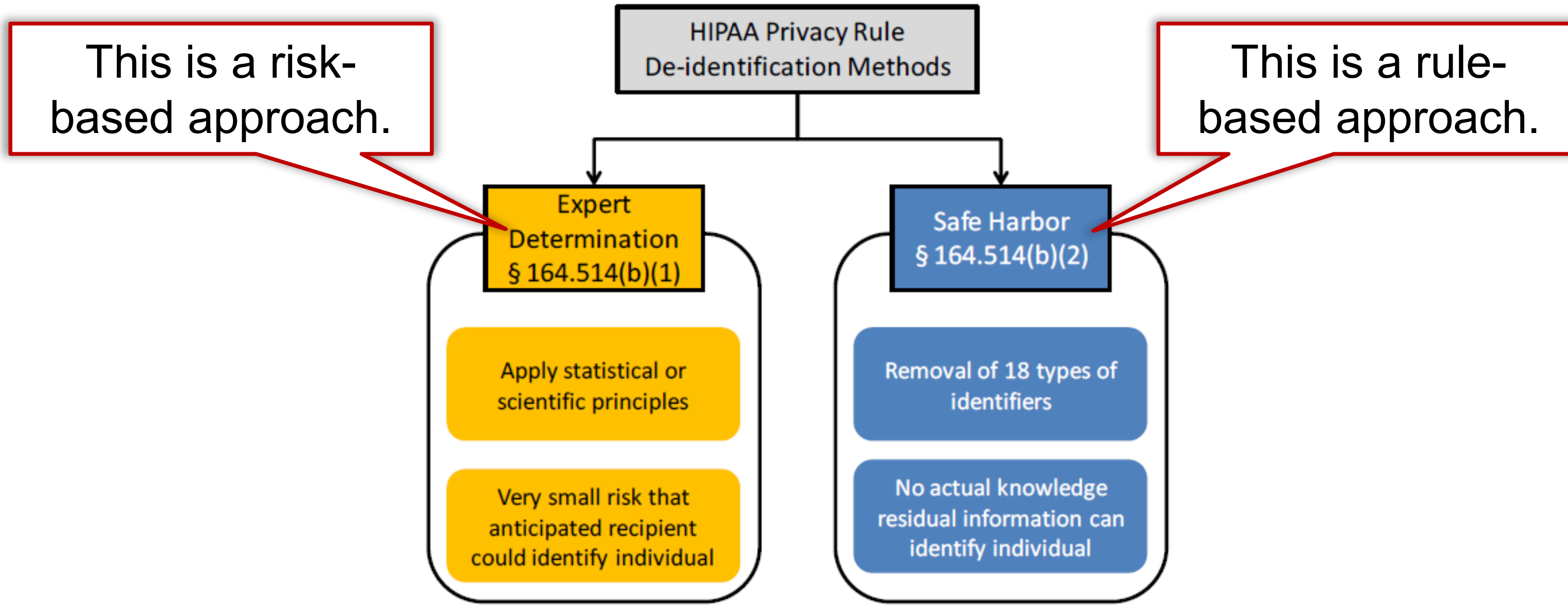
The Term *De-Identification* Under the US HIPAA



HIPAA Privacy Rule is split into two methods:

1. Safe Harbor – standardized way of de-identifying data
2. Expert Determination – an expert reviews the dataset for re-identification vulnerability

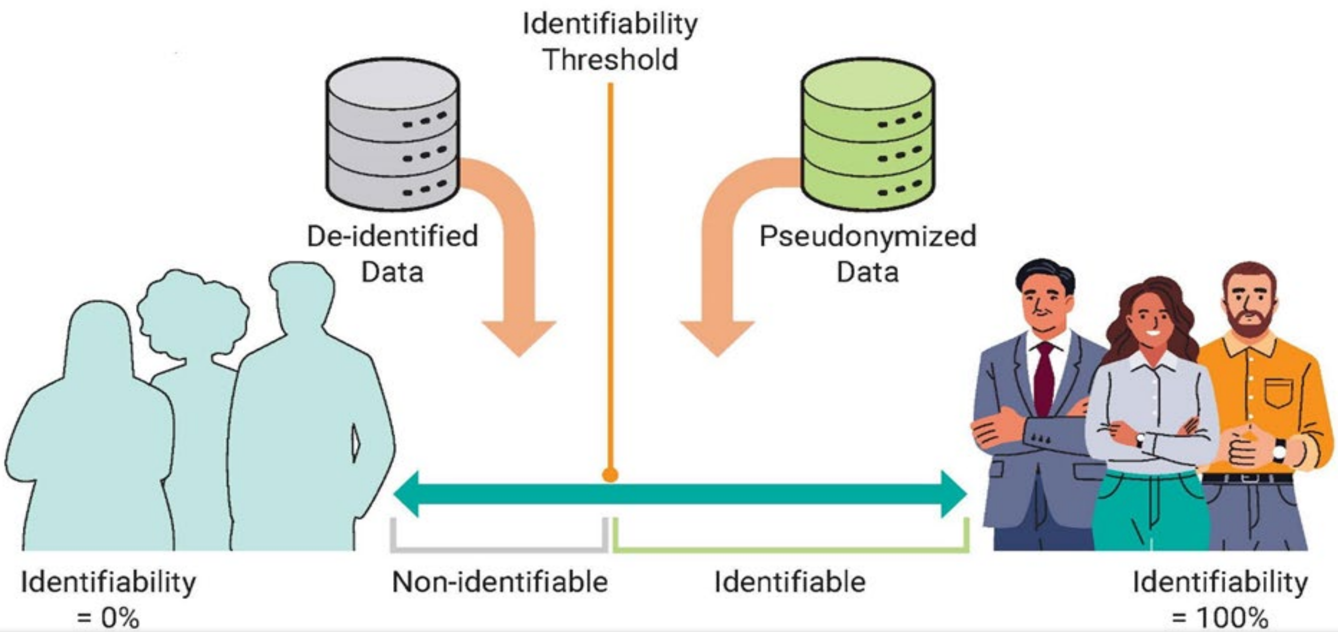
The Term *De-Identification* Under the US HIPAA



HIPAA Privacy Rule is split into two methods:

1. Safe Harbor – standardized way of de-identifying data
2. Expert Determination – an expert reviews the dataset for re-identification vulnerability

Our Terminology on the Spectrum



A properly de-identified dataset no longer contains information that identifies an individual or information that could be used, either alone or with other information, to identify an individual based on what is reasonably foreseeable in the circumstance.

image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

Terminology and Scope

SETTING THRESHOLDS

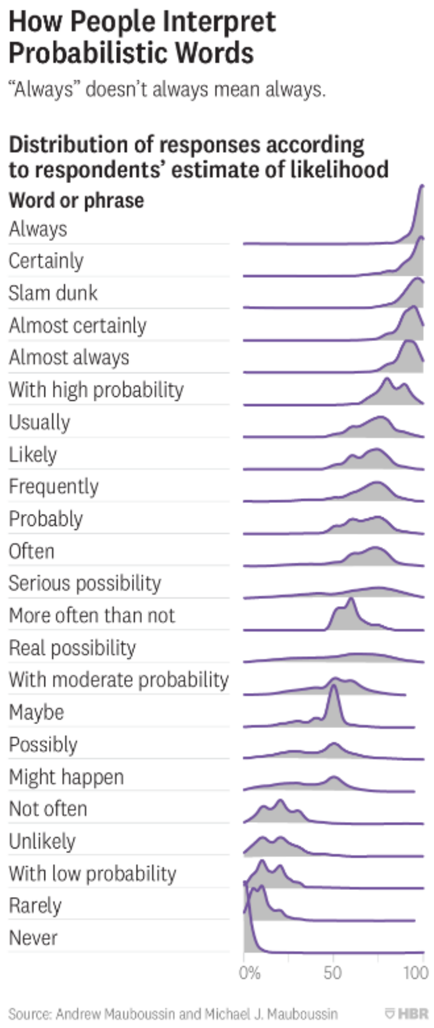
Setting Thresholds

Expressing acceptable risk qualitatively — using words like "unlikely" or "very small" — introduces significant ambiguity. Research shows that different individuals assign widely varying numeric probabilities to the same verbal descriptors, with meaningful gender differences in interpretation.

For defensible de-identification practice, **quantitative thresholds** are required. Fortunately, a rich body of precedents from national statistical agencies, large data custodians, regulators, ISO standards, and court cases provides a reliable basis for selecting appropriate numeric thresholds.

uOttawa

What are acceptable risk thresholds?



Acceptable risk is often expressed verbally or qualitatively; but what do these subjective expressions of probability really mean ?

EA. Mauboussin and M. J. Mauboussin, "If You Say Something Is 'Likely,' How Likely Do People Think It Is?," *Harvard Business Review*, Jul. 03, 2018.

uOttawa

Standards & Guidelines as Precedents

ISO/IEC 27559: 2022

Scenario	Content (Matrix)	Threshold
Public	High possibility of attack, low impact	Max 0,1
	High possibility of attack, medium impact	Max 0,075
	High possibility of attack, high impact	Max 0,05
Non-public	Low-med possibility of attack, low-medium impact	Avg 0,1
	Medium possibility of attack, medium impact	Avg 0,075
	Medium-high possibility of attack, medium-high impact	Avg 0,05

Ontario De-identification Guidance, 2025

Invasion of Privacy Values	Re-identification Risk Threshold (very low)	Cell Size Equivalent
Low	0.09	11
Medium	0.075	15
High	0.05	20

This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

uOttawa

Precedent Risk Thresholds

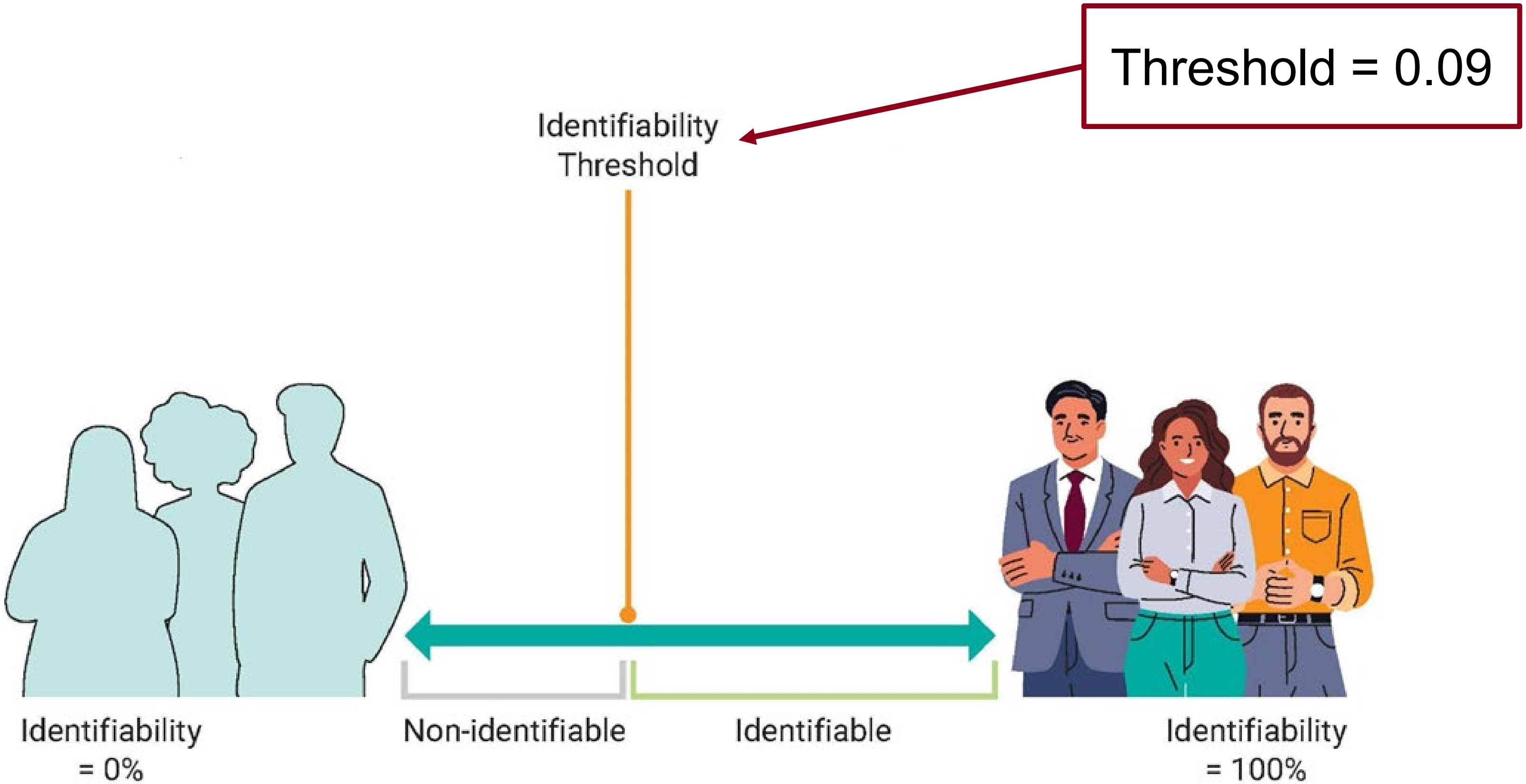


image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

uOttawa

Practical De-Identification Methods

CONCEPTS IN RISK-BASED DE-IDENTIFICATION

uOttawa

Module Agenda

1. **What is risk** — Defining the components of re-identification risk
2. **The adversary** — Re-identification threats and attacks
3. **Disclosure types** — Understanding the type of information disclosure that occurs
4. **Risk-based de-identification methodology** — Overview of the risk-based de-identification methodology

uOttawa

Concepts in Risk-Based De-Identification

WHAT IS RISK

Components of Re-Identification Risk

Vulnerability of the Data × Probability of Attack

Controls to Manage the Probability of Attempt (i.e., Context Risk)

Vulnerability of the Data

×

Probability of Attack

↑

Security controls

Privacy controls

Contractual controls

uOttawa

Multiple Levers to Manage Risk

Vulnerability of the Data × Probability of Attack

Vulnerability of the Data:

- "Traditional" De-Identification
- Synthetic Data Generation

Probability of Attack:

- Security controls
- Privacy controls
- Contractual controls

Privacy-Utility Trade-Off

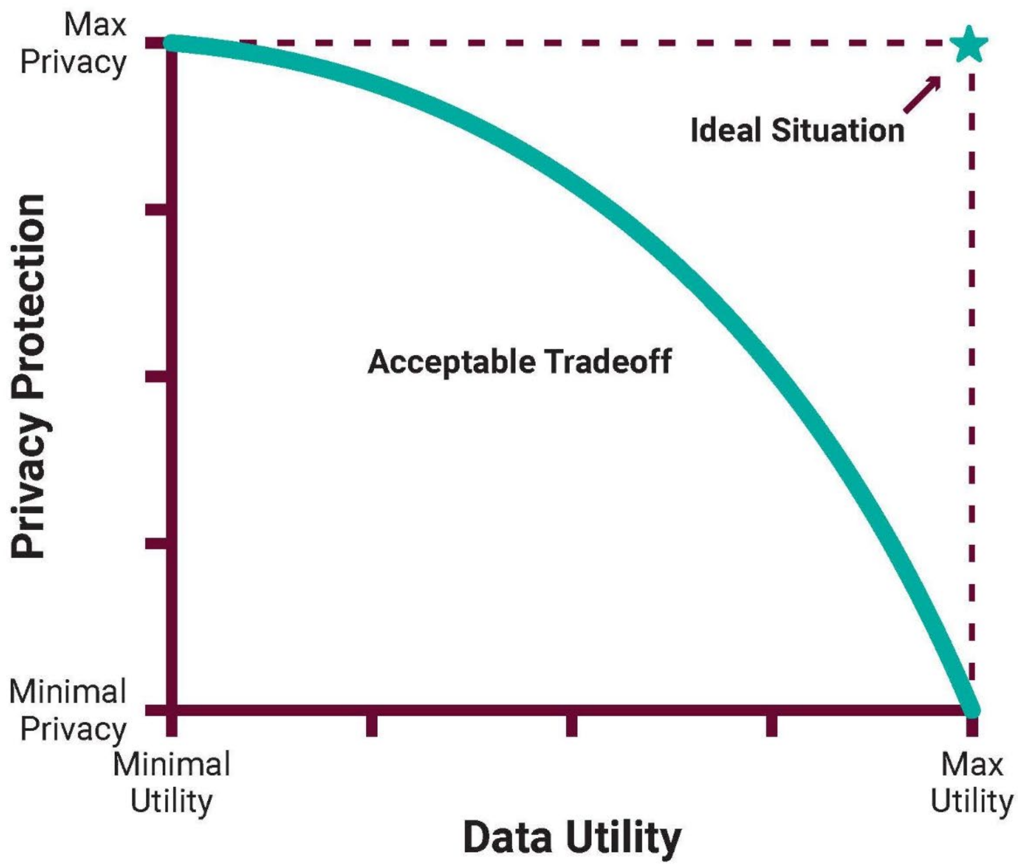


image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

Concepts in Risk-Based De-Identification

THE ADVERSARY

The Adversary in Non-Public and Public Release Scenarios

- In a non-public release scenario, threats arise primarily from the **anticipated recipient**.

- In a public release scenario, the data is accessible to unrestricted actors, and a deliberate attack is the dominant model.

uOttawa

Three Types of Attacks

- The term *attack* refers broadly to three types of threats

  1. Deliberate attacks
  2. Inadvertent recognition
  3. Data breaches

- These three attacks provide reasonable coverage of plausible attacks – managing these would provide defensible protection.

uOttawa

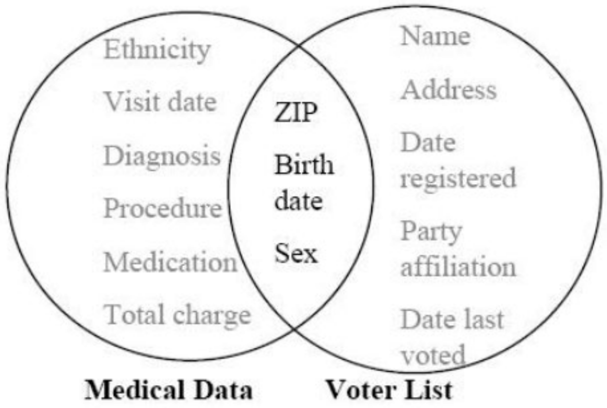
A Deliberate Adversary – Who Are They?

- This is a generic term intended to indicate the individual or entity that may attempt to re-identify or attack a dataset to cause a disclosure to occur.
- It is not intended to be derogatory term – we just need to call them something.
- Other terms commonly used: attacker, intruder, snooper.
- Can be, for example, a neighbor, relative, co-worker, media, academic.
- Most published attacks on data have been performed by the media and academics.

K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A Systematic Review of Re-identification Attacks on Health Data," *PLoS ONE*, vol. 6, no. 12, 2011. doi:10.1371/journal.pone.0028071
Meurers T, Baum L, Haber AC, et al. Health Data Re-Identification: Assessing Adversaries and Potential Harms. *Stud Health Technol Inform*. 2024;316:1199–203. doi: 10.3233/SHTI240626

The Adversary – A Researcher

- The Massachusetts Group Insurance Commission (GIC) released health records of state employees under the assumption that it was de-identified.
- A researcher, L. Sweeney, used that data and conducted a linking attack with the Massachusetts voter registry.
- She linked the two datasets using ZIP, birth date and sex.
- And successfully re-identified Governor William Weld's medical record, including diagnoses and medications.
- In fact, 87%* of Americans could be singled out using only ZIP code, birth date, and gender.



From: L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

*According to P. Golle (Revisiting the uniqueness of simple demographics in the US population. Proceedings of the 5th ACM Workshop on Privacy in Electronic Society. 2006), it is rather at 63% of the US population

Background Knowledge of An Adversary

- Any re-identification attempt requires the availability of background knowledge.
- This comes from
  - using public sources (e.g., registries, media, professional organizations, employers).
  - due to access to non-public sources of information (e.g., dataset from other research projects).
  - being an acquaintance (e.g., neighbor, co-worker) of the target individual with access to private background information about the individual.
- Quasi-identifiers operationalize this concept and represent the variables an adversary may know and use for re-identification.

Concepts in Risk-Based De-Identification

DISCLOSURE TYPES

Components of Re-Identification Risk

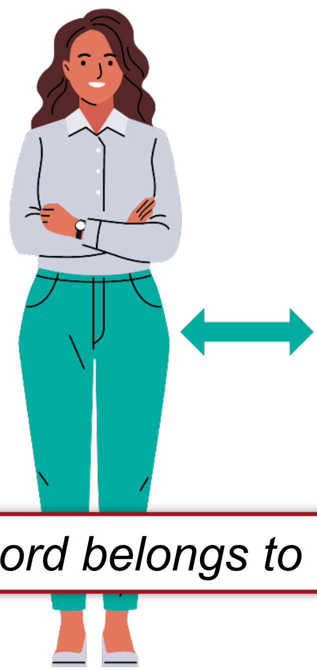
Vulnerability of the Data × Probability of Attack

Data vulnerability can be assessed for one (or multiple) disclosure type(s).

uOttawa

Identity disclosure is when a person's identity is assigned to a record

Sally



Indirect Identifiers		New Information
Sex	**Year of Birth**	**Grade [%]**
Male	1992	60
Male	1999	77
Male	1994	89
Female	2001	56
Female	**1997**	**84**
Male	2000	68
Male	2002	95
Female	1996	83
[illegible]	1993	79
[illegible]	2003	91
Male	1995	66

Which record belongs to Sally?

image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

uOttawa

Types of Disclosures

- Identity disclosure
- Attribute disclosure
- Membership disclosure

These are types of inferences.

In this workshop, we are going to focus on **identity disclosure**.

But we will provide a conceptual perspective on **attribute** and **membership disclosure** as well.

uOttawa

Attribute disclosure is when personal information is inferred from attributes without identifying an individual's record

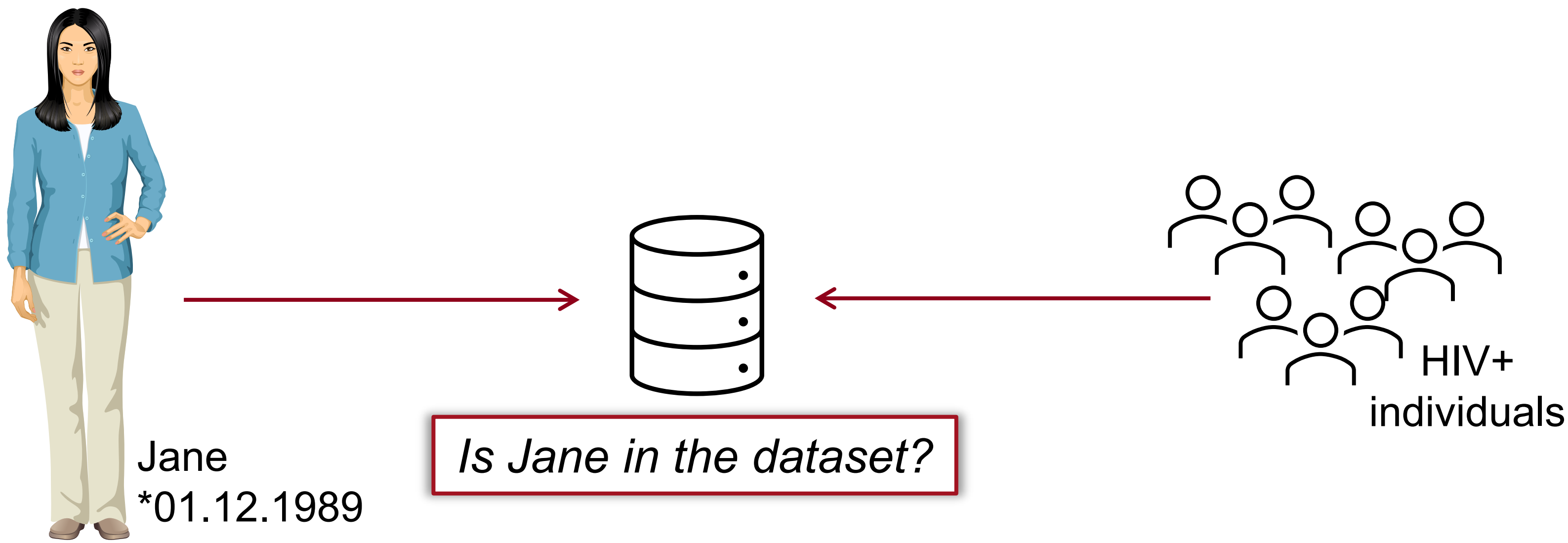
Jane
*01.12.1989

Sex	Year of Birth	Diagnosis
Female	**1989**	**Breast Cancer**
Female	1952	Breast Cancer
Female	1985	Cervical Cancer
Female	1989	Breast Cancer
Female	1989	Breast Cancer

What can we learn about Jane's diagnosis?

uOttawa

Membership disclosure is when a person’s membership in a dataset is inferred



Focus on Identity Disclosure Risk

- In this workshop, identifiability refers to the risk of identity disclosure.
- An identifiability threshold therefore corresponds to a threshold on identity disclosure risk.

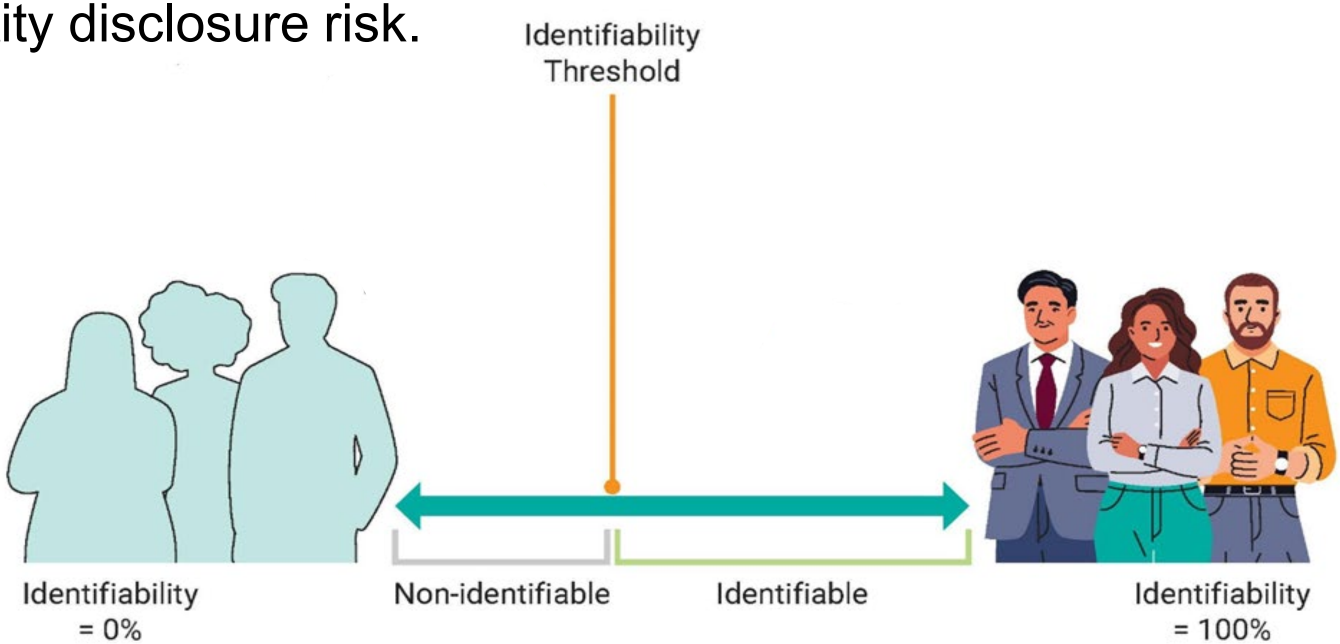
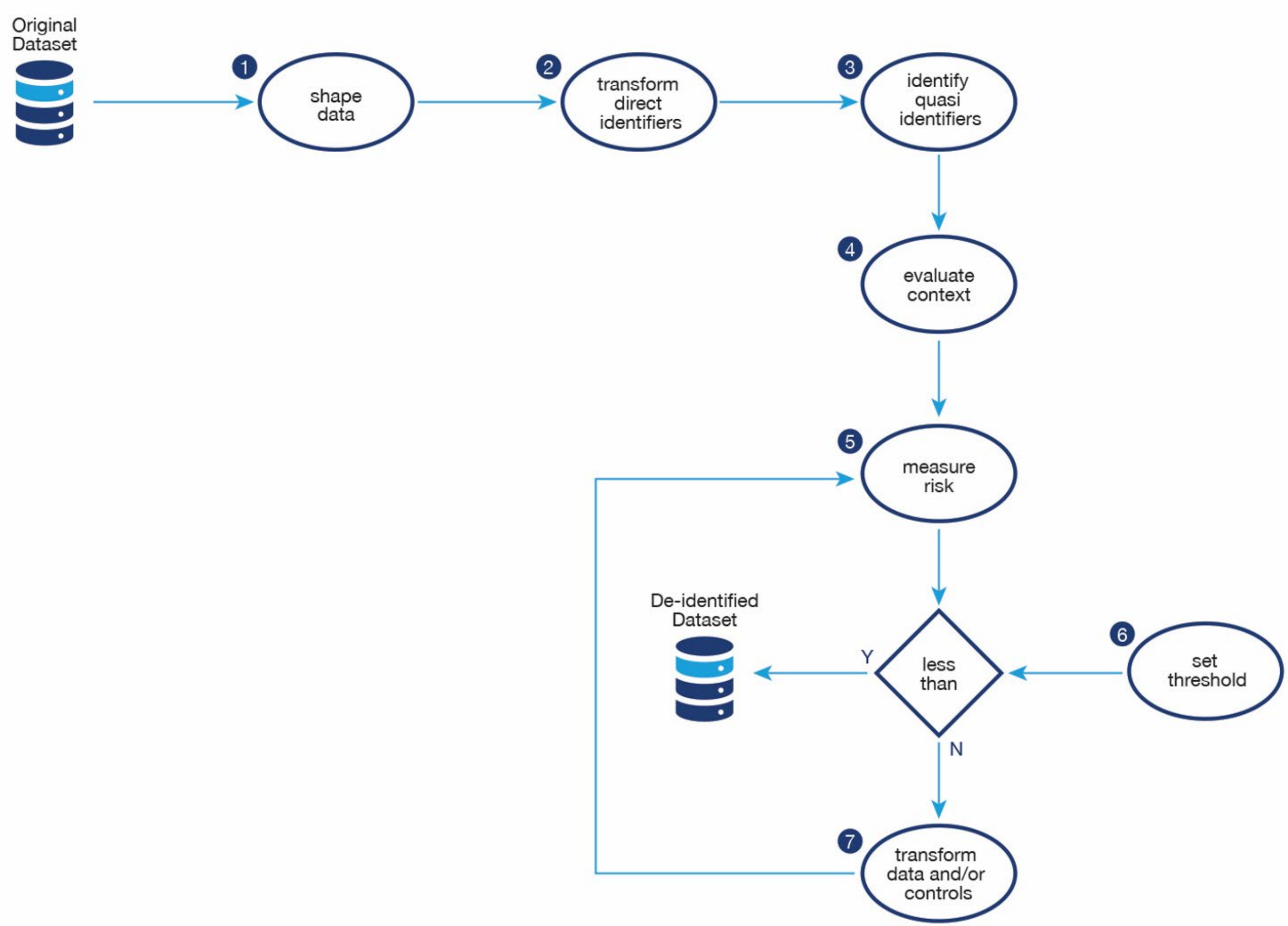


image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

Concepts in Risk-Based De-Identification

RISK-BASED DE-IDENTIFICATION METHODOLOGY

Risk-Based De-Identification



Demonstration 1

AUTOMATED RISK ASSESSMENT

Automated Risk Assessment

Goal: Measuring risk in a (de-identified) COVID-19 dataset

Materials:

- COVID-19 data
  - original version: artificially created dataset with high identifiability
  - de-identified version: transformed data via generalization
- Information about the population where the dataset is drawn from
- EviData Tool by Woodway Assurance*

* Woodway Assurance is a spin-off from our lab that provides automated, independent third-party risk assessment of de-identified, anonymized and synthetic data. https://www.woodway-assurance.com/

Practical De-Identification Methods

MEASURING VULNERABILITY

Module Agenda

1. **Variable Classification**: Identifying direct identifiers, quasi-identifiers and sensitive variables
2. **Measuring identity disclosure vulnerability**: Understanding the basics of vulnerability assessment
3. **Identity disclosure metrics**: Details of how to measure average disclosure vulnerability

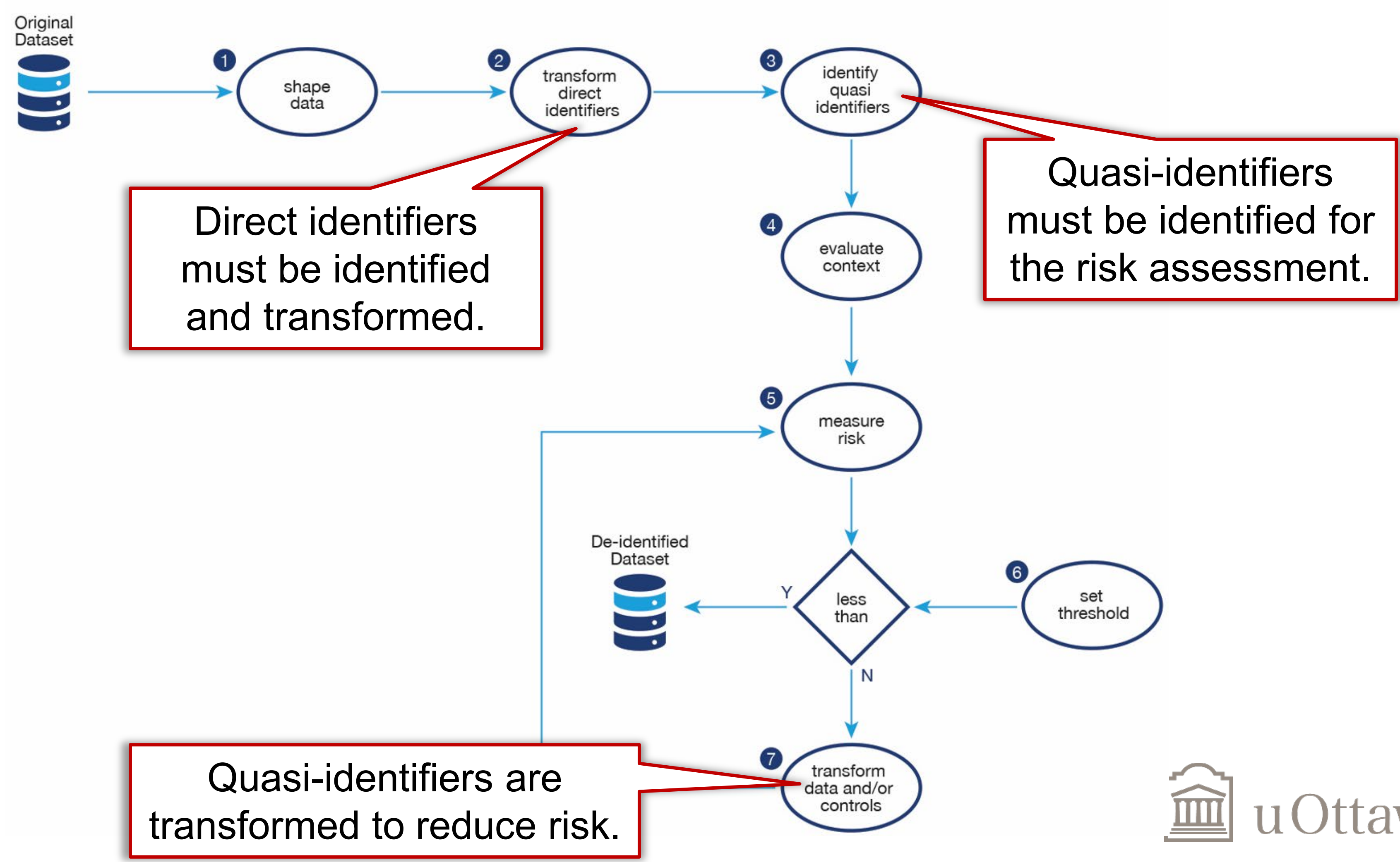
uOttawa

Measuring Vulnerability

VARIABLE CLASSIFICATION

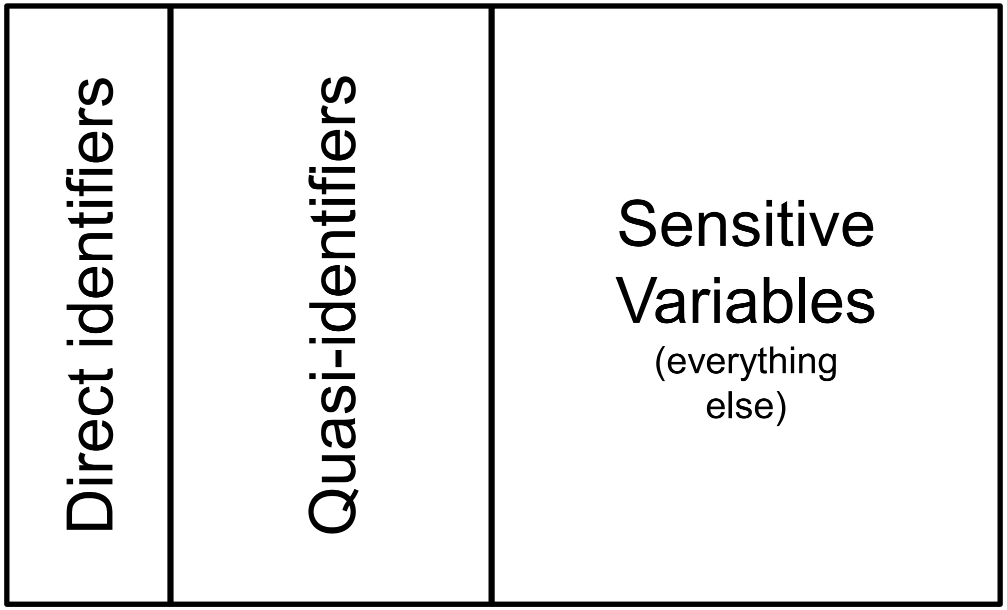
uOttawa

Where Variable Classification Matters for De-Identification



uOttawa

Classifying Variables in a Dataset



This classification is important to determine how they will be treated during the vulnerability assessment and the data transformation process.

uOttawa

Direct Identifiers

- These are variables in a dataset that can directly and uniquely identify an individual.
- Typical examples are a SIN / SSN, and health insurance number.
- Names are also considered direct identifiers.
- Direct identifiers are pseudonymized (i.e., removed or replaced to eliminate direct identifiability).
- In most cases, this would not affect the analytic utility of a dataset.

Examples of Direct Identifiers

- names;
- street addresses (other than town, city, province and postal code);
- telephone numbers;
- fax numbers;
- e-mail addresses;
- Social Insurance Numbers;
- medical records numbers;
- health insurance numbers;
- full face photos
- account numbers;
- certificate license numbers;
- vehicle identifiers and serial numbers, including license plates;
- device identifiers and serial numbers;
- URLs;
- IP address numbers;
- biometric identifiers (including finger and voice prints);

Direct and Indirect Identifiers

DIRECT IDENTIFIERS

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

pseudonymization

QUASI-IDENTIFIERS

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

uOttawa

Pseudonymous Data is Typically Still Personal Information

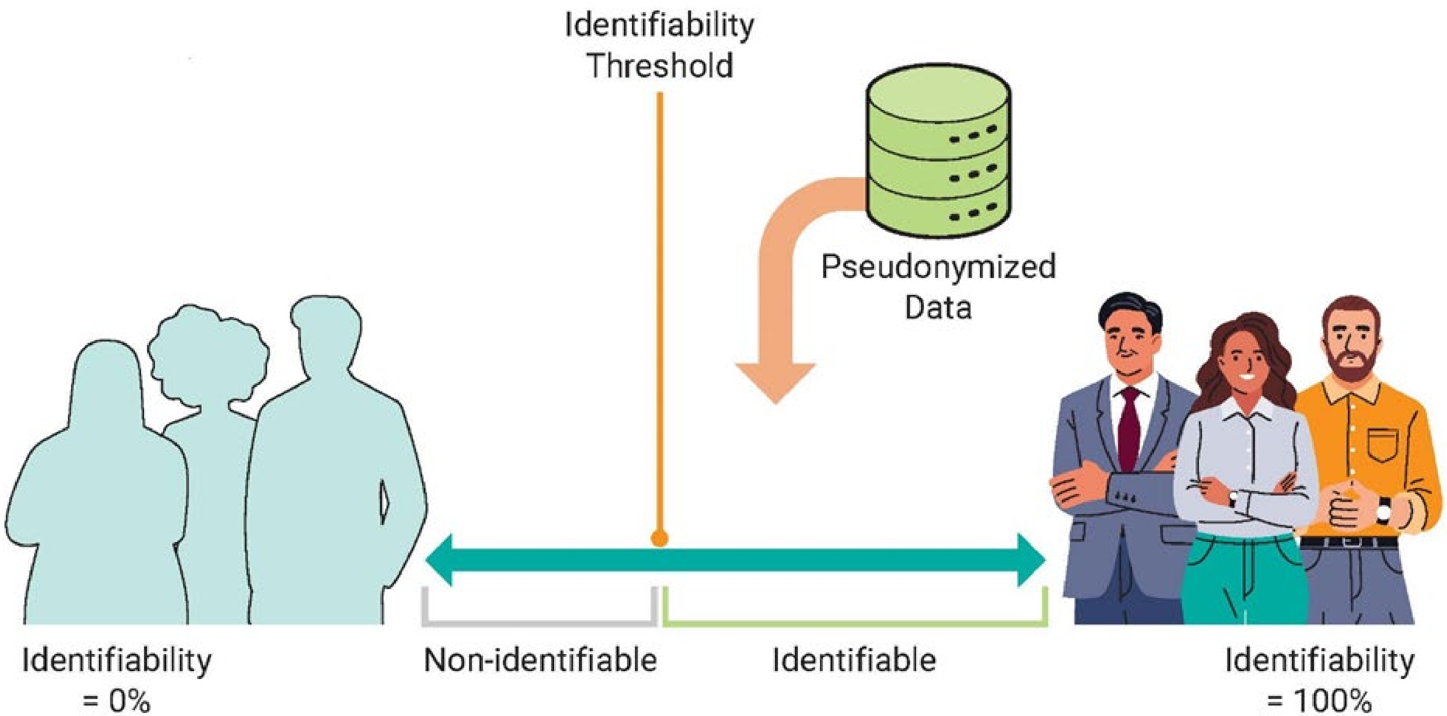


image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

Quasi-Identifiers (or Indirect Identifiers)

- After pseudonymization, risk is driven primarily by quasi-identifiers.
- Quasi-identifiers represent variables that can be known by an adversary and be used, in combination, to identify an individual.
- Risk modeling requires explicit assumptions about the adversary knowledge.
- Many variables could, in principle, be known by some adversary.
- However, it is very unlikely that **any single adversary** possesses knowledge of all such variables.

The concept of *adversary power* formalizes assumptions about the scope of adversary background knowledge. It specifies the maximum number of quasi-identifiers (e.g., 7) plausibly simultaneously known.

Direct and Quasi-Identifiers

DIRECT IDENTIFIERS

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

QUASI-IDENTIFIERS

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

de-identification

Measuring Vulnerability

MEASURING IDENTITY DISCLOSURE VULNERABILITY

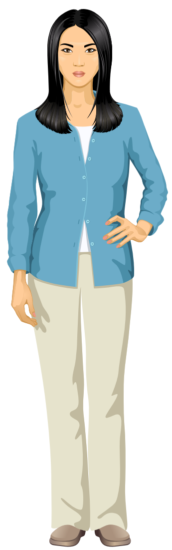
Components of Re-Identification Risk

Vulnerability of the Data × Probability of Attack

We measure identity disclosure vulnerability.

uOttawa

Identity disclosure is when a person's identity is assigned to a record



Quasi-identifier | Sensitive variable

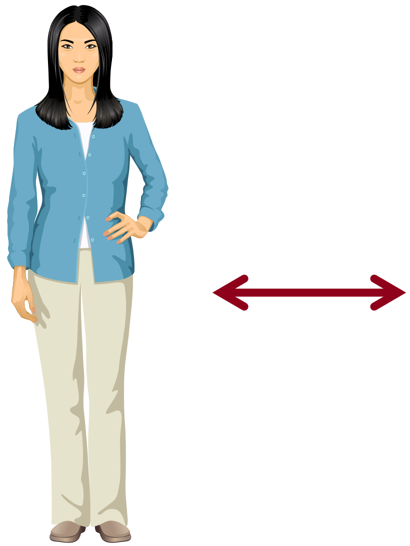
Sex	Year of Birth	NDC
Male	1975	009-0031
Male	1988	0023-3670
Male	1972	0074-5182
Female	1993	0078-0379
Female	**1989**	**65862-403**
Male	1991	55714-4446
Male	1992	55714-4402
Female	1987	55566-2110
Male	1971	55289-324
Female	1996	54868-6348
Male	1980	53808-0540

One record matches to one individual using quasi-identifiers.

uOttawa

Generalization means that more than one record can match a person

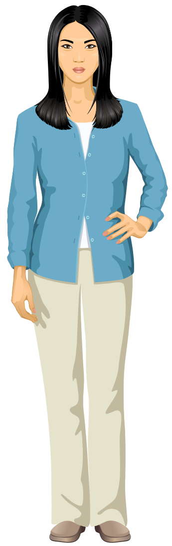
Generalization means that more than one record can match a person



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

Generalizing the *Year of Birth* increases the number of records that can match one individual.

Vulnerability is measured by the equivalence class size

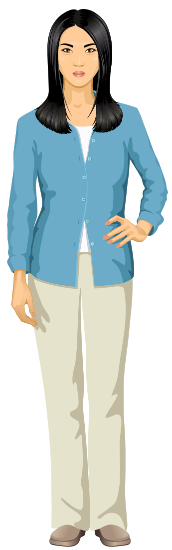


Sex	Year of Birth	NDC	Class Size	Vulnerability
Male	1975	009-0031	1	1
Male	1988	0023-3670	1	1
Male	1972	0074-5182	1	1
Female	1993	0078-0379	1	1
Female	**1989**	**65862-403**	**1**	**1**
Male	1991	55714-4446	1	1
Male	1992	55714-4402	1	1
Female	1987	55566-2110	1	1
Male	1971	55289-324	1	1
Female	1996	54868-6348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

One record matches to one real person will have the highest vulnerability = 1.

uOttawa

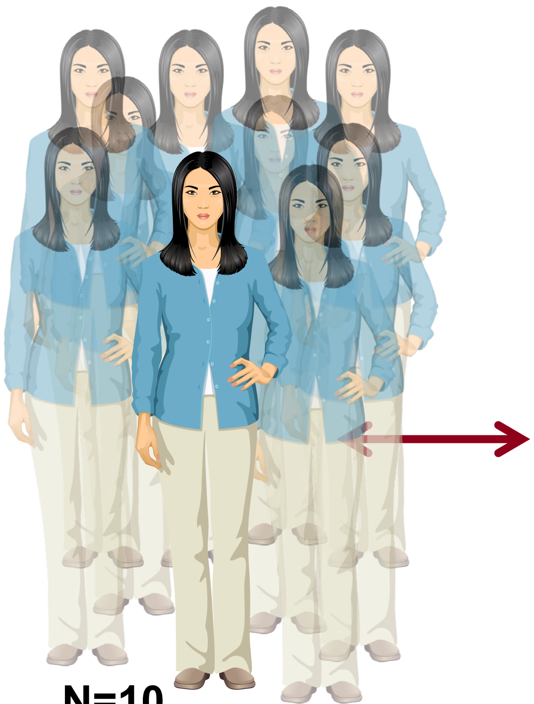
When we generalize the class size gets bigger, so the vulnerability decreases



Sex	Year of Birth	NDC	Class Size	Vulnerability
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

Generalizing the data, increasing the number of possible record matches (class size) from 1 to 2 matches, decreases the vulnerability by 50% (vulnerability = 0.5).

But the population (class) size also matters



N=10

Sex	Year of Birth	NDC	Sample Class Size	Population Class Size
Male	1970-1979	009-0031	3	15
Male	1980-1989	0023-3670	2	10
Male	1970-1979	0074-5182	3	15
Female	1990-1999	0078-0379	2	10
Female	**1980-1989**	**65862-403**	**2**	**10**
Male	1990-1999	55714-4446	2	10
Male	1990-1999	55714-4402	2	10
Female	1980-1989	55566-2110	**2**	**10**
Male	1970-1979	55289-324	3	15
Female	1990-1999	54868-6348	2	10
Male	1980-1989	53808-0540	2	10

Matching one data record to one individual has a vulnerability = 1. If the population size increases, to where one data record can match 10 individuals in the population, the vulnerability decreases = 1/10 = 0.1.

Measuring Vulnerability

IDENTITY DISCLOSURE METRICS

Calculating Identity Disclosure Vulnerability

There are different ways to evaluate the identity disclosure vulnerability from a given dataset:

- Average vulnerability
- Maximum vulnerability
- Uniqueness vulnerability
- Strict average vulnerability

Average Vulnerability	Maximum Vulnerability	Uniqueness Vulnerability	Strict Average Vulnerability

This workshop will focus on one (average vulnerability) to exemplify the calculations. In real world, different metrics would be applied depending on the use case.

A Simplistic Example Dataset

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

sample →

Name	Sex	Year of Birth	NDC
Sarah Petrova	Female	1993	0078-0379
Emily Ndlovu	Female	1989	65862-403
Daniel Kowalski	Male	1991	55714-4446
Isabella Nguyen	Female	1989	54868-6348

↓ pseudonymize

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

This example dataset is a sample from a larger population and is intended to be shared. We must assess the identity disclosure vulnerability.

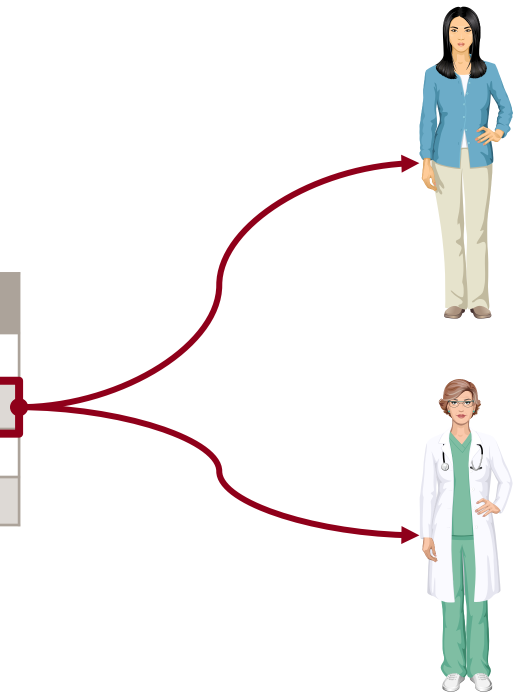
Quasi-identifiers: sex, year of birth

Sensitive variable: NDC (national drug code)

uOttawa

Attacks can be in two directions – sample to population attack (s2p)

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348



s2p example: trying to link one record in the dataset to a real person.

uOttawa

Average Vulnerability: Sample to Population Attack (s2p)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Sample to Population Attack (s2p)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

Average Vulnerability: Sample to Population Attack (s2p)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Sample to Population Attack (s2p)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

Average Vulnerability: Calculate s2p vulnerability

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/1

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Calculate s2p vulnerability

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/3

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Calculate s2p vulnerability

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/2

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Sample to Population Attack (s2p)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/3

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Average s2p Vulnerability

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Average vulnerability:

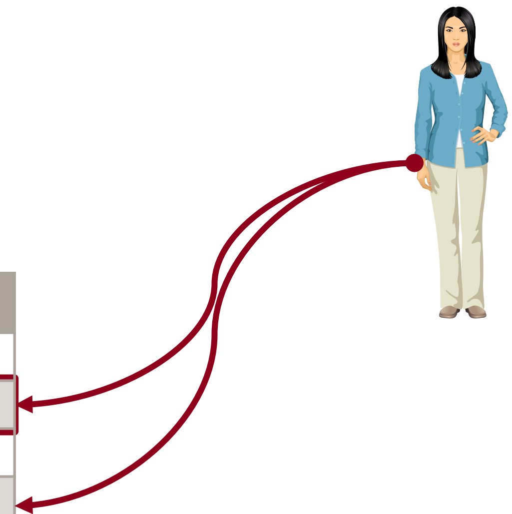
$$1/4 \times (1/1 + 1/3 + 1/2 + 1/3) = 0.54$$

Step 3: average the vulnerability across all records in the microdata.

uOttawa

Attacks can be in two directions – population to sample attack (p2s)

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348



p2s example: trying to identify an individual in the dataset.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare microdata (i.e., sample) records to population using indirect identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

not part of the microdata

Step 1: compare sample records to population using quasi-identifiers.

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Step 1: compare sample records to population using quasi-identifiers.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 0

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/1

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/2

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/1

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Vulnerability: 1/2

Step 2: calculate vulnerability as 1 divided by the number of records that match in the population.

uOttawa

Average Vulnerability: Population to Sample Attack (p2s)

Population

Name	Sex	Year of Birth
James Smith	Male	1975
Michael Tanaka	Male	1988
David Martínez	Male	1992
Sarah Petrova	Female	1993
Emily Ndlovu	Female	1989
James Haddad	Male	1991
John Okoro	Male	1992
Olivia Novak	Female	1989
Daniel Kowalski	Male	1991
Sophia Diop	Female	1996
Matthew Sørensen	Male	1980
Isabella Nguyen	Female	1989

Sample

Sex	Year of Birth	NDC
Female	1993	0078-0379
Female	1989	65862-403
Male	1991	55714-4446
Female	1989	54868-6348

Average vulnerability:

$$1/12 \times (1/1 + 1/2 + 1/1 + 1/2) = 0.25$$

Step 3: average the vulnerability across all target individuals in the population.

uOttawa

Average Vulnerability

- Average vulnerability is the maximum of the average p2s and s2p vulnerability.
- In the example, it is the maximum out of 0.25 (p2s) and 0.54 (s2p).
- Information about the population is necessary to calculate average vulnerability.
  - If not available, this information can be estimated using published population estimators*.
  - Using information from the sample only does not reflect an accurate vulnerability in most cases.
  - Using information from the sample only can be correct if the sample is equal to the population or the adversary knows which targets are in the data.

* Jiang Y, Mosquera L, Jiang B, Kong L, El Emam K. Measuring re-identification risk using a synthetic estimator to enable data sharing. *PLoS One*. 2022;17(6):e0269097. doi:10.1371/journal.pone.0269097).

uOttawa

Practical De-Identification Methods

MODELING THE ATTACKS

Module Agenda

Probability of a deliberate attack — 1 — Modeling the probability of a deliberate attack

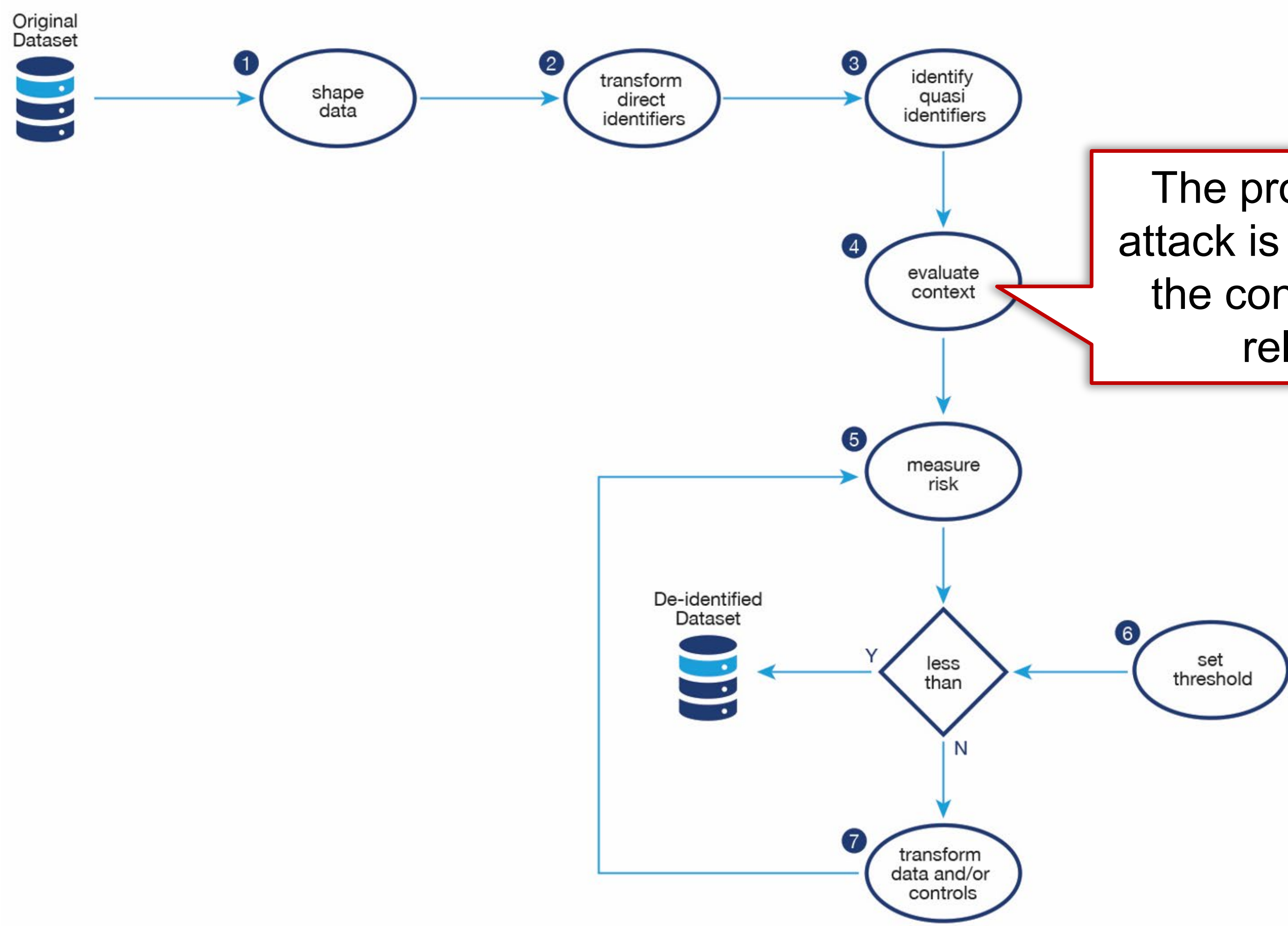
Inadvertent recognition — 2 — Understanding the probability of an inadvertent recognition

Data breaches — 3 — How likely is a data breach

Measuring risk — 4 — Modeling risk based on vulnerability and probability of attack

uOttawa

The Context of a Data Release Informs the Probability of Attack



The probability of attack is informed by the context of the release.

uOttawa

Components of Re-Identification Risk

Vulnerability of the Data × Probability of Attack

We consider deliberate attacks, inadvertent recognitions and data breaches.

Modeling the Attacks

PROBABILITY OF A DELIBERATE ATTACK

uOttawa

A Deliberate Attack

- A deliberate attack is performed by an adversary with intent.
- The adversary must
  - have access to the dataset.
  - be motivated and have the capacity to re-identify the dataset.
- The probability of an attack therefore depend on:
  - the security, privacy, and contractual controls in place at the data recipients.
  - the motives and capacity of the data recipient to attempt a re-identification attack.

Assessing Probability of A Deliberate Attack

Privacy, Security, and Contractual Controls	Motives and Capacity	Probability of Re-Identification Attack
High	Low	0.15
	Medium	0.2
	High	0.25
Medium	Low	0.25
	Medium	0.3
	High	0.4
Low	Low	0.4
	Medium	0.5
	High	0.6

Most organizations have MEDIUM controls and MEDIUM motive and capacity for an attempt.

A checklist for mitigating controls can be found in the 2025 updated IPC guidelines. We do not cover them in this workshop.

image:
This has been reproduced from the *De-identification Guidelines for Structured Data* published by the Office of the Information and Privacy Commissioner of Ontario. 2025.

uOttawa

Modeling the Attacks

INADVERTENT RECOGNITION

uOttawa

An Inadvertent Attack

- An inadvertent attack occurs when an individual working with the dataset inadvertently recognizes someone that they know.
- The recognized individual may be a neighbor, colleague or relative.
- Recognition is more likely when the data recipient operates in the same geographic context as the data subjects.
- The direction attack is typically p2s (population-to-sample).
- The probability of inadvertent recognition can be modeled as a function of overlap between the recipient population and the data subjects.

$$1-(1-p)^{m}$$

p = Population overlap
m = Number of acquaintances

uOttawa

Knowing Someone With Breast Cancer

- 80,486 female individuals were diagnosed with breast cancer in the previous 10 years according to a CCO report form 2024.
- Combined with the size of the female population at that time, this gives a prevalence of 1%.
- The number of acquaintances can be informed by Dunbar's number (i.e., 150). For female friends, m can be set to 75.

The likelihood of having a female acquaintance with breast cancer in Ontario would then be **53%**.

uOttawa

Modeling the Attacks

DATA BREACHES

uOttawa

Data Breaches in the Healthcare Sector

- The probability of a data breach can be reduced by implementing better security, privacy, and contractual controls.
- We can use published numbers about the frequency of data breaches to estimate the likelihood of a breach occurring.
- We want to err on the conservative side while doing so.

Numbers from the 2022 U.S. Department of Health and Human Services HIPAA compliance report suggest a yearly probability of 0.092 (0.126 adjusted for underreporting)*.

* A value of 0.27 has been proposed in El Emam, Khaled. Guide to the de-identification of personal health information. CRC Press, 2013 based on historical numbers.

uOttawa

Modeling the attack

MEASURING RISK

Components of Re-Identification Risk

Vulnerability of the Data × Probability of Attack

Deliberate Attack Model

Identity Disclosure Vulnerability

×

Probability of a Deliberate Attack

Identity Disclosure Vulnerability → Average vulnerability: max(s2p,p2s)

Probability of a Deliberate Attack → MEDIUM controls and MEDIUM motives and capacity

0.54 x 0.3 = 0.16

Inadvertent Attack Model

Identity Disclosure Vulnerability

×

Probability of Spontaneous Recognition

Average vulnerability: p2s

Probability of knowing someone with breast cancer

0.25 x 0.53 = 0.13

Breach Attack Model

Identity Disclosure Vulnerability

×

Probability of a Data Breach

Average vulnerability:

max(s2p,p2s)

Estimated probability of a breach in the healthcare sector

0.54 x 0.126 = 0.07

uOttawa

The Overall Risk Value

To err on the conservative side, we take the max. value out of the three attack models…

$$\max \begin{pmatrix} \text{deliberate attack} \\ \text{inadvertent attack} \\ \text{data breach} \end{pmatrix} = \max \begin{pmatrix} 0.16 \\ 0.13 \\ 0.07 \end{pmatrix} = 0.16$$

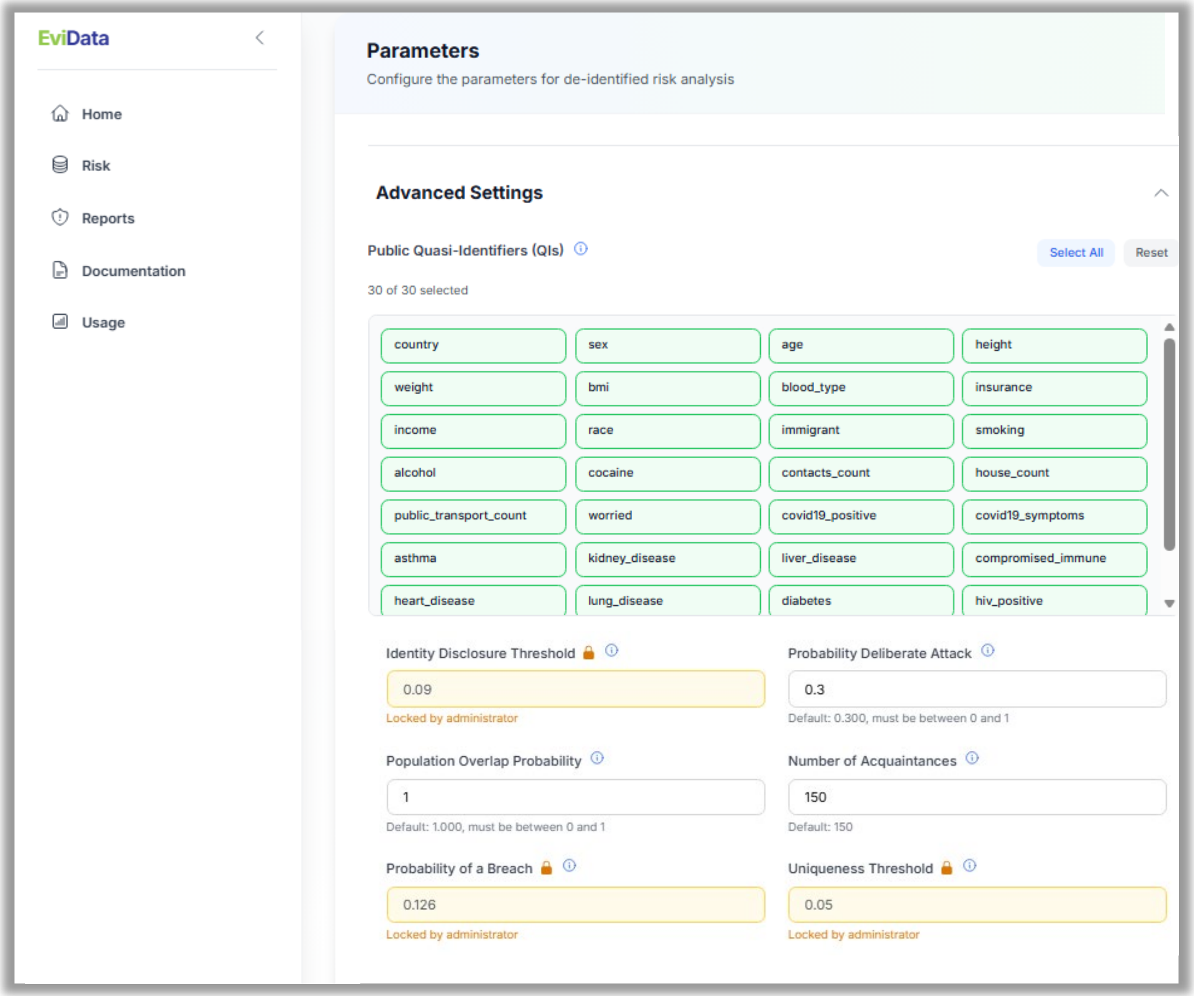
uOttawa

Demonstration 2

UNDERSTANDING THE RISK REPORT

uOttawa

Changing Assumptions in Automated Risk Assessment



- Default values have been established based on literature.
- Parameters can still be adjusted where appropriate.
- The (estimated) size of the population is the only parameter that cannot easily be automated.

uOttawa

Automated Risk Assessment – Output

Independent assessment using EviData(TM)
Certified by Woodway Assurance
Date: February 21, 2026 at 17:58 UTC
Report ID: WWA-D-000002

EviData™
POWERED BY Woodway Assurance

De-identified Data Evaluation Report

Data Custodian: Lisa Pilgram (lpilgram@ehealthinformatio...
Data Recipient: Anticipated Recipient(s)
Date: February 21, 2026 at 12:58 PM (EST)
Report ID: WWA-D-000002
File name: nexoid_course_sample.csv
Generated by: EviData™ by Woodway Assurance Ltd (version 1.5.0, contact: info@woodway-assurance.com, www.woodway-assurance.com)

The Identity Disclosure Risk is HIGH.

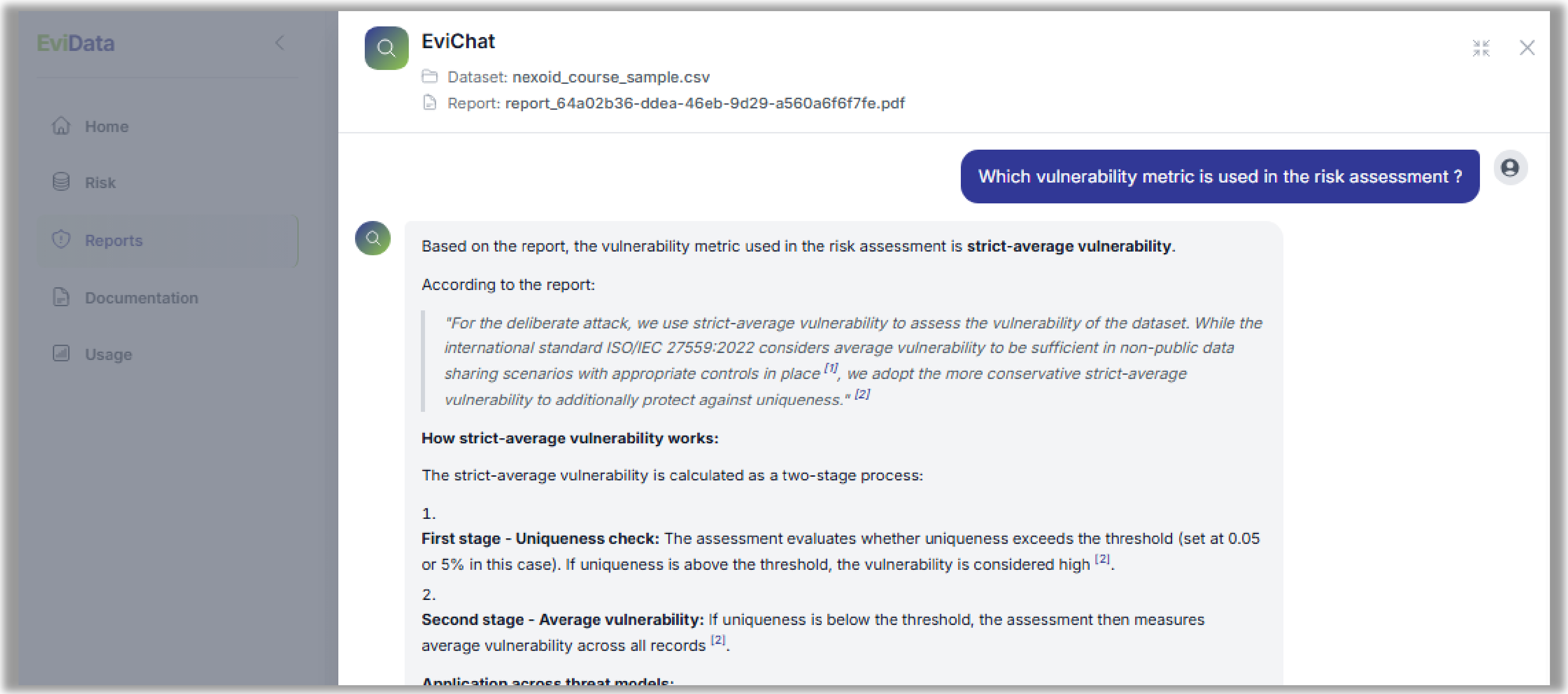
Disclaimer: This assessment is based on information provided by the Data Custodian, along with assumptions made in the context of the assessment. The Data Custodian is responsible for verifying that these assumptions remain valid for their use case. A re-assessment is required if conditions or assumptions change in a material way. Results are valid for up to **2 years** from the above date or before if any condition or assumption is no longer met.

The output of EviData is a detailed De-Identified Data Evaluation Report. It includes:

- the final estimated identity disclosure risk
- the vulnerabilities and risks for all three attack types
- the assumptions that were taken in the context of the assessment
- a brief description of the dataset under evaluation
- the detailed risk measurement methodology

uOttawa

Leveraging AI Companions to Understand the Report



Khaled El Emam, PhD
Canada Research Chair in Medical AI, University of Ottawa
Senior Scientist, Children's Hospital of Eastern Ontario Research Institute
Professor, School of Epidemiology and Public Health, University of Ottawa
kelemam@ehealthinformation.ca

Lisa Pilgram, MD
Postdoctoral Fellow, University of Ottawa
Clinician Scientist, Department of Nephrology and Medical Intensive Care, Charité - Universitaetsmedizin Berlin
lpilgram@ehealthinformation.ca

Thank you!