What is ‘reasonable’? Exploring when a de-identification process is sufficiently protective

Vance Lockton
Senior Technology and Policy Advisor

Information and Privacy Commissioner of Ontario
Commissaire à l’information et à la protection de la vie privée de l’Ontario

NRC Responsible Data Speaker Series

February 1, 2022

Key Takeaways

- Privacy law generally relies on conceptual (rather than technical) definitions of de-identification
- Documentation of your processes is critical to proving ‘reasonableness’
- There is a lot of room, and desire, for standards

A sidebar on the current political environment

Factors to keep in mind

- Era of privacy legislation reform in Canada
  - This will have to consider the fundamental question of “what data is covered by the law”

- I argue there is a significant tension growing between “supporting innovators” and concern about uses of de-identified information
  - On latter, see:
    - Germany Google Fonts finding
    - Austria (and EDPS, and potentially Dutch) Google Analytics finding
    - Reaction to PHAC

Reasonable / Sufficiently Protective De-Identification

Reasonable / sufficient for what?

De-identification as a safeguard

vs.

De-identification to remove data from scope of privacy law

Pseudonymization

GDPR Definition

- "… personal data [that] can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures [to prevent re-identification]"

Benefits

- Taken into consideration when determining purpose consistency, data protection by design, safeguards, etc.

Cont'd

Pseudonymization

Bill 64: "For the purposes of this Act, personal information is de-identified if it no longer allows the person concerned to be directly identified;"

Ontario Privacy Reform White Paper: "de-identified information" means information about an individual that no longer allows the individual to be directly or indirectly identified without the use of additional information.

(Note: a factor in determining whether purposes are appropriate, whether info can be used for internal research and development)

BC PIPA Reform Report: Recommendation 3 – "Ensure that PIPA include definitions of pseudonymized information as personal information, and anonymized information as outside the scope of PIPA, similar to definitions in the GDPR. "

Anonymization

PIPEDA / Privacy Act

> *"personal information* means information about an identifiable individual …"
>
> (see also OPC Interpretation Bulletin on Personal Information)

Ontario Personal Health Information Protection Act (PHIPA)

> "personal health information … means identifying information about an individual …"

Gordon v. Canada: Information will be about an "identifiable individual" where there is a serious possibility that an individual could be identified through the use of that information, alone or in combination with other information

Anonymization

Ontario Private Sector White Paper:

"… information [that] has been altered irreversibly, according to generally accepted best practices, in such a way that no individual could be identified from the information, whether directly or indirectly by any means or by any person.

Quebec Bill 64:

- "information … is anonymized if it irreversibly no longer allows the person to be identified directly or indirectly. Information anonymized under this Act must be anonymized according to generally accepted best practices."

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by
FUTURE OF PRIVACY FORUM
FPF.org

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

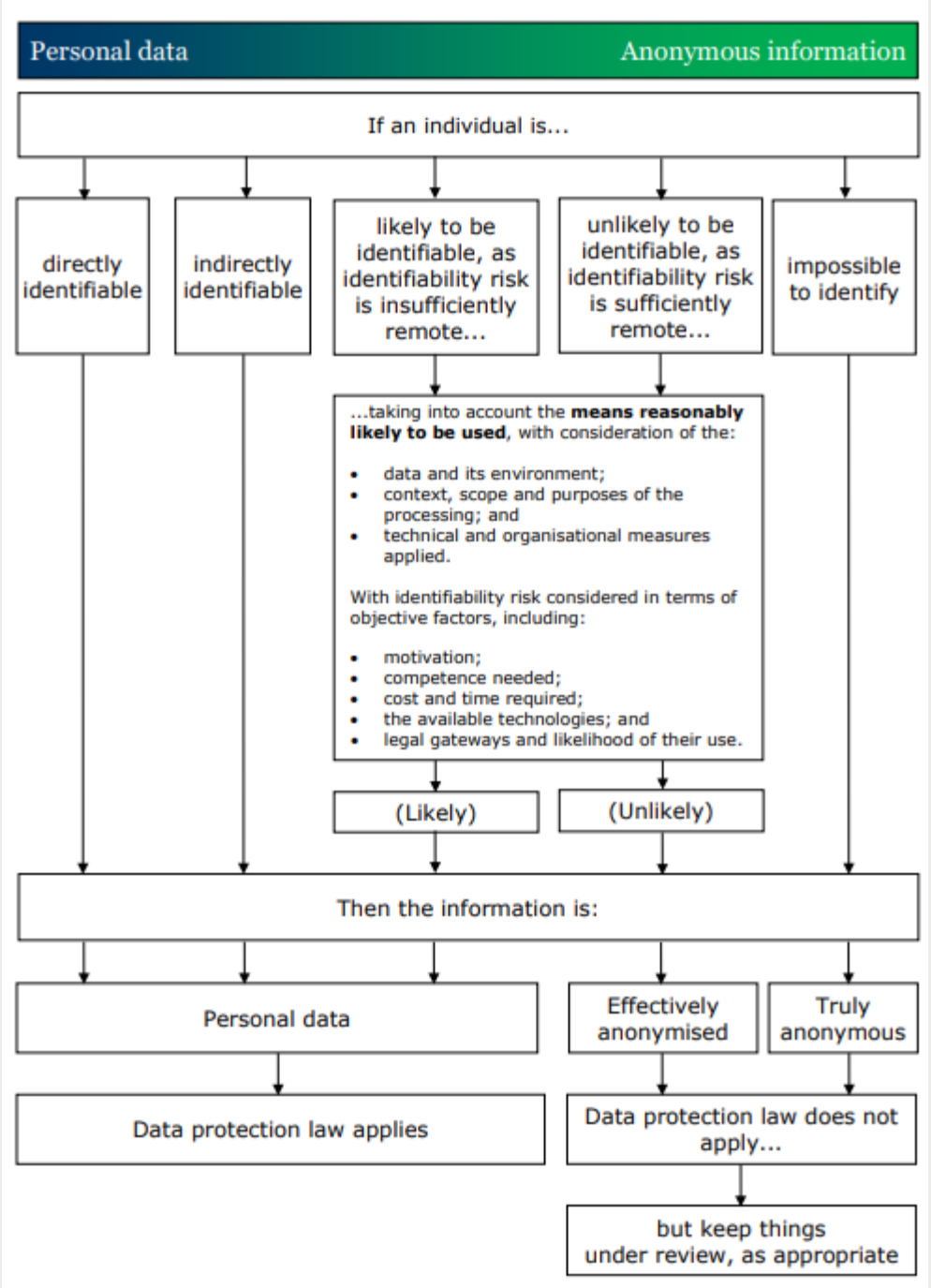
DEGREES OF IDENTIFIABILITY
Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA
Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA
Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA
Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)



Via: UK Information Commissioner's Office

Draft anonymization, pseudonymization and privacy enhancing technologies guidance (October 2021)

Chapter 2: How do we ensure anonymization is effective?

Regulators’ Approach

Why should we accept your assessment?

For example:

- Have you followed guidance from a regulator?
  - See, for instance, IPC's De-Identification Guidelines for Structured Data
- Have you followed a peer-reviewed de-identification process?
- Have you followed an recognized standard or code of practice?
- Have you measured re-identification risk level? How does it compare to industry/sectoral norms?
- Has an independent entity provided a re-identification risk assessment?

AND MOST IMPORTANTLY

- Can you prove it?

Underlying assumptions

- What threat model needs to be considered?
  - “Motivated intruder” vs. state actor?
  - Available data vs. accessible data

- What does “re-identification” mean?
  - Is it enough to re-identify a single person in the dataset?
  - Is it enough to infer information about an individual based on the anonymized dataset?

IPC Approach to De-Identification

- De-Identification guidelines developed in partnership with Khaled El Emam
  - Winner of the inaugural International Conference of Data Protection and Privacy Commissioners' (ICDPPC) award for excellence in research.

- Current (and future) approach has connections to federal OPC; Canadian Anonymization Network; similarities to UK Information Commissioner's anonymisation code of practice

- De-identification: the process of removing any information that (i) identifies an individual, or (ii) for which there is a reasonable expectation that the information could be used, either alone or with other information, to identify an individual.

Nine-Step Process

1. Determine the release model
2. Classify variables
3. Determine an acceptable re-identification risk threshold
4. Measure the data risk
5. Measure the context risk
6. Calculate the overall risk
7. De-identify the data
8. Assess data utility
9. Document the process

Risk Threshold

Invasion of Privacy	Re-identification Risk Threshold	Cell Size Equivalent
Low	0.1	10
Medium	0.075	15
High	0.05	20

- "Invasion of privacy" is a factor of:
  - Sensitivity; level of detail; number of individuals; individuals' expectations; etc.

Data Risk

- Regulators will be reasonably ambivalent about process used – I leave most of this discussion to other experts

- Fundamentally – likelihood that an attack will succeed.

Context Risk

- Non-public releases:
  - Deliberate insider attack
    - Extent of controls set out in data sharing agreement
    - Motives and capacity of the recipient

Privacy and Security Controls	Motives and Capacity	Probability of Re-identification Attack
High	Low	0.05
	Medium	0.1
	High	0.2
Medium	Low	0.2
	Medium	0.3
	High	0.4
Low	Low	0.4
	Medium	0.5
	High	0.6

Context Risk (Cont'd)

- Non-public releases (cont'd)
  - Inadvertent recognition
  - Data breach
- Semi-public or public release
  - Assume re-identification attack will occur, unless clear reason not to.

- Fundamentally, likelihood that attack will occur.

Overall Risk

$$P(\text{Re-Id}) = P(\text{Re-Id} \mid \text{Attack}) * P(\text{Attack})$$

(Step 10) Governance

- Protection against attribute disclosure
- On-going and regular re-identification risk assessments
- Audit data recipients to ensure contractual compliance
- Examine disclosure of overlapping data
- Accountability

Reasonableness beyond Reasonable Protection

- De-identification does not necessarily give license for all uses of data; see, for instance, PIPEDA s.5(3)

- At minimum, transparency is a critical element
  - See Cadillac Fairview or PHAC location monitoring for what can happen when people are surprised

- Open questions on limits
  - See article on use of suicide hotline texts to develop customer service AI
    - ("Suicide hotline shares data with for-profit spinoff, raising ethical questions" Politico, Jan. 28, 2022)

What’s Next

Potential future developments

- Canadian Code of Practice for De-Identification?
  - Via CIOSC? CANON? Other?

- Ability for regulators to formally review / approve codes of practice?

- Changes in definitions and/or how de-identified information can be used? Shift away from identifiability as determiner of applicability of privacy law?
  - Report on the India ~~Personal~~ Data Protection Bill:
    - "It is impossible to distinguish between personal data and non-personal data, when mass data is collected or transported."

Discussion Questions

Discussion Questions

- Does the “three-state” model of identifiability allow for socially-beneficial research / uses of data? If not, what would you change?

- What information and/or guidance could regulators provide to support privacy-respectful use of data?

- How can we advance a standard or code of practice for de-identification?

Questions for the IPC?

- Happy to discuss now, or feel free to reach out at:

Vance.Lockton@ipc.on.ca or info@ipc.on.ca

HOW TO CONTACT US

Information and Privacy Commissioner of Ontario

2 Bloor Street East, Suite 1400

Toronto, Ontario, Canada M4W 1A8

Phone: (416) 326-3333 / 1-800-387-0073

TDD/TTY: 416-325-7539

Web: www.ipc.on.ca

E-mail: info@ipc.on.ca

Media: media@ipc.on.ca / 416-326-3965