Identifiability Definitions

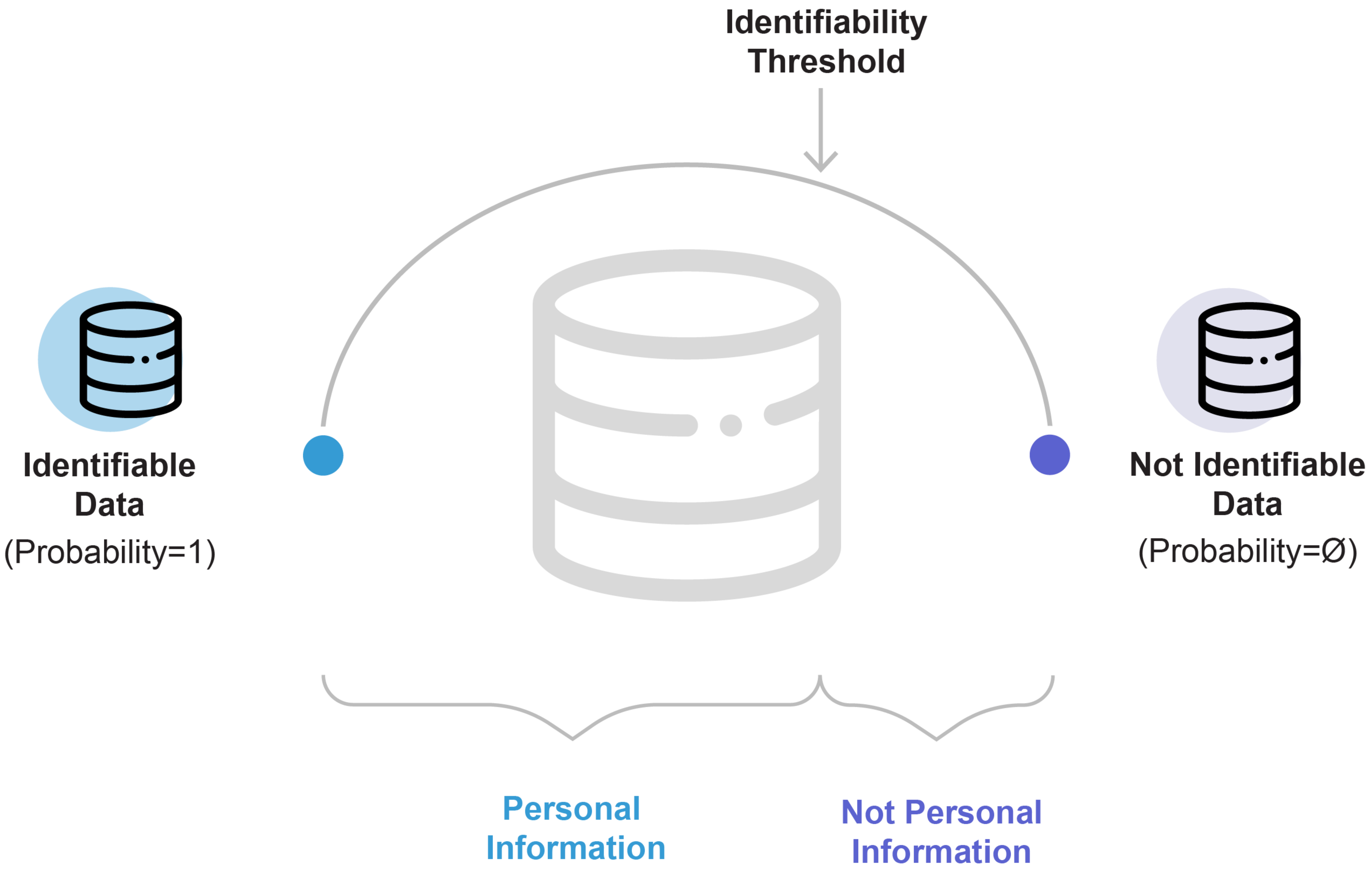
Khaled El Emam

21st July, 2022

Replica Analytics

AN AETION COMPANY

Identifiability spectrum and risk thresholds



Identifiability Threshold

Identifiable Data

(Probability=1)

Not Identifiable Data

(Probability=Ø)

Personal Information

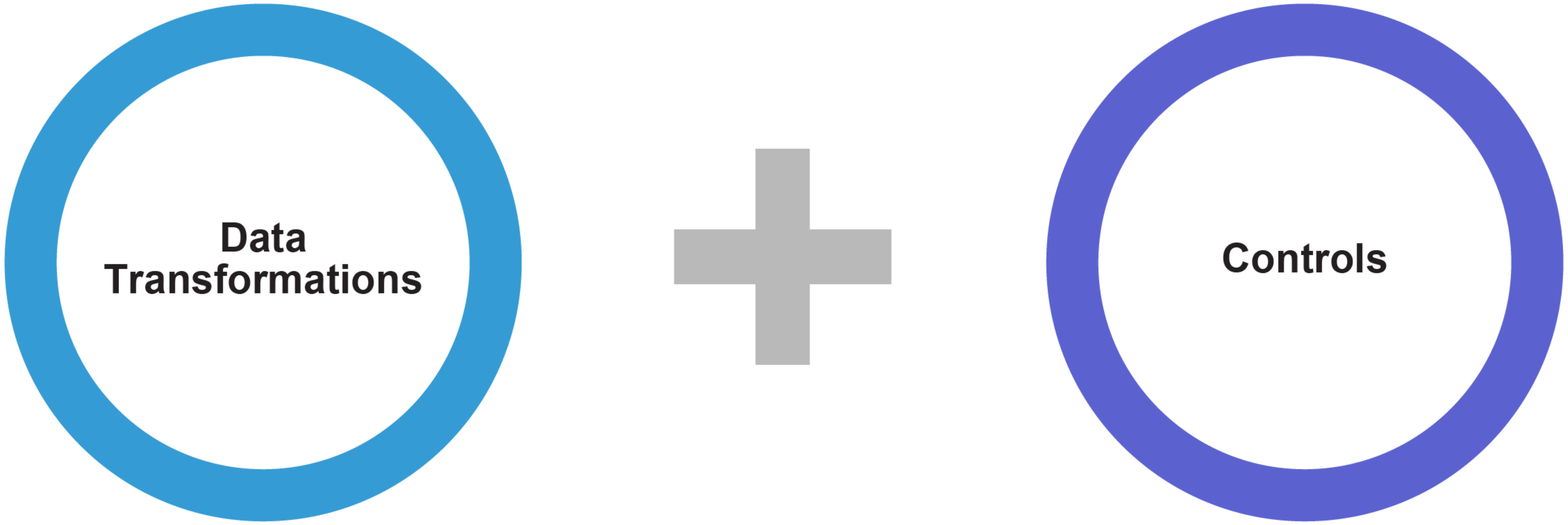
Not Personal Information

(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics

AN AETION COMPANY

Risk takes into account the controls in place at the data recipient



Data Transformations

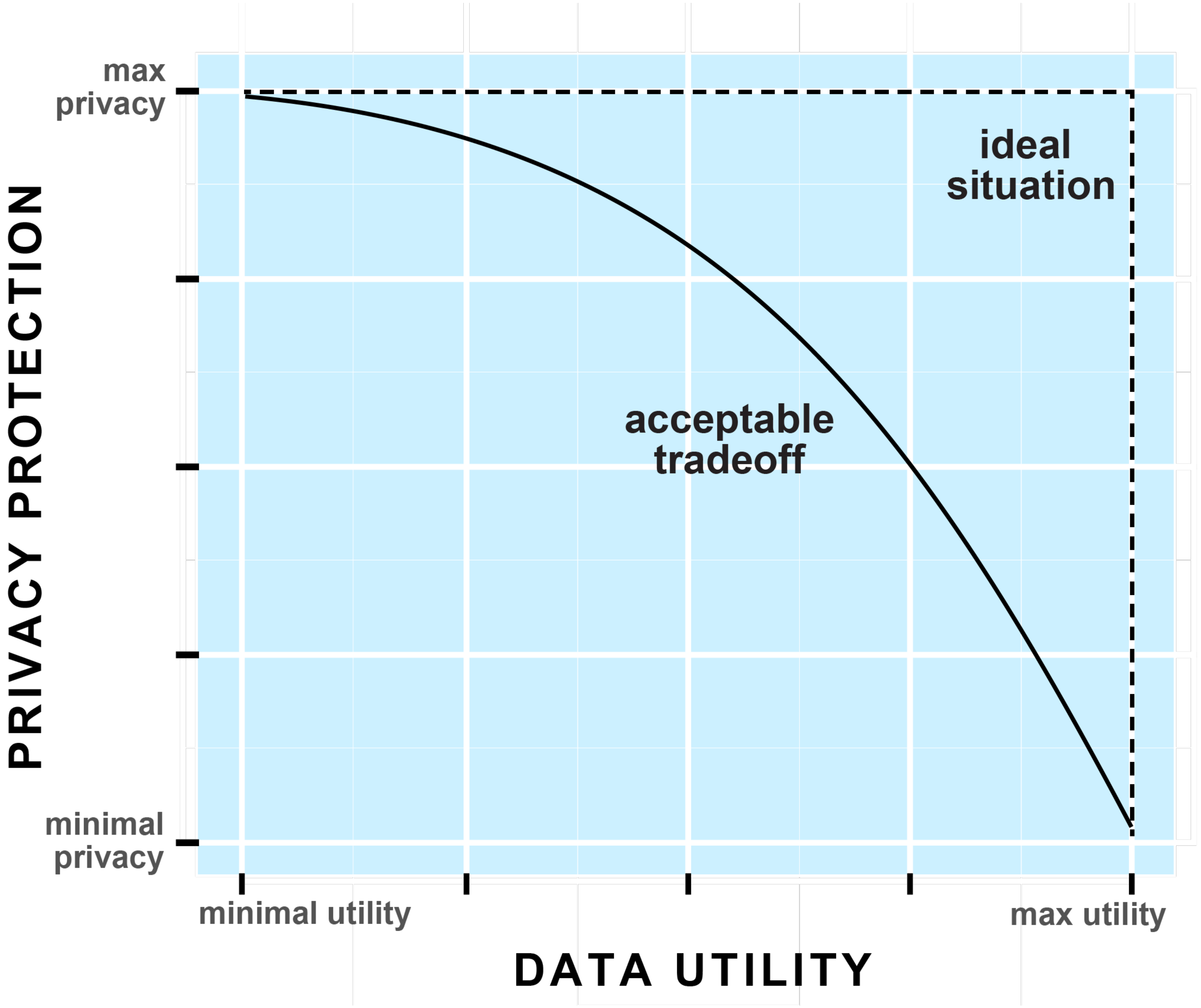
Controls

- Encryption / hashing / tokenization
- Generalization & Suppression
- Addition of noise
- Synthetic data generation

- Security controls
- Privacy controls
- Contractual controls

(c) Copyright 2019-2022 Replica Analytics Ltd.

Privacy-Utility Trade-off



Replica Analytics
AN AETION COMPANY

(c) Copyright 2019-2022 Replica Analytics Ltd.

The focus is on identity disclosure, but it is not the only type of data risk

- Identity disclosure
- Attribution disclosure
- Membership disclosure

(c) Copyright 2019-2022 Replica Analytics Ltd.

Exploring the Concept of Identifiability in Canada's Proposed *Consumer Privacy Protection Act* (Bill C-27)

Adam Kardash

Chair, Privacy and Data Management

Osler, Hoskin & Harcourt LLP

National Lead, AccessPrivacy

OSLER

Background

- Bill C-27 tabled in Canadian Parliament in June 2022.
- Bill C-27 proposes:
  - The *Consumer Privacy Protection Act* ("CPPA") – a private sector privacy statute that would replace the privacy regime in Part I of PIPEDA.
  - The *Artificial Intelligence and Data Act* ("AIDA") – a legislative framework that would regulate the creation and use of artificial intelligence systems.
- The CPPA and AIDA contain several specific concepts related to the concept of "identifiability".

Key CPPA Terms Relating to the Concept of "Identifiability"

- Under the CPPA, the concept of "personal information" refers to "information about an identifiable individual"(s. 2(1)).

- Supreme Court of Canada jurisprudence has established a very broad scope of this definition of "personal information" (*Dagg v Canada (Minister of Finance),* [1997] 2 SCR 403; *Canada (Information Commissioner) v. Canada (Commissioner of the RCMP),* 2003 SCC 8; *Canada (Information Commissioner) v. Canada (Transportation Accident Investigation and Safety Board),* 2006 FCA 157).

- In essence, Canadian jurisprudence provides that information will be deemed about an "identifiable individual" where it is "serious possibility" for an individual to be identified through the use of that information, alone or in combination with other available information. (*Gordon v. Canada (Minister of Health),* 2008 FC 258; see also Canada *(Information Commissioner) v. Canada (Transportation Accident Investigation and Safety Board)*, 2006 FCA 157; see also OPC *Interpretation Bulletin: Personal Information* (2013)).

Key CPPA Terms Relating to the Concept of “Identifiability” (Cont’d.)

- Under the CPPA, the concept of personal information includes:
  - Sensitive personal information
    - personal information of minors is expressly deemed to be sensitive (s. 2(2))
    - Personal information that is contextually sensitive
  - Personal information in “de-identified” form

CPPA Concept of "De-identified" Personal Information

- Under the CPPA, de-identified personal information is:
  - Personally identifiable information that has been modified such "that an individual cannot be directly identified from it, though a risk of the individual being identified remains" (s. 2(1)).
- Generally, personal information that has been de-identified is deemed by the CPPA to be personal information, except in connection with the following statutory provisions:
  - certain exceptions to consent
  - requests for disposal
  - access requests
  - mobility rights
  - de-identification prohibitions
  - retention

  (See sections 20, 21, 22(1), 39(1), 55, 56, 63(1), 71, 72, 74, 75 and 116.) (s. 2(3)).

Bill C-27 - Concept of "Anonymized" Data

- CPPA clarifies that the "Act does not apply to personal information that has been anonymized" (Section 6(5)).
- Under the CPPA, the concept of "anonymized" means:
  - to irreversibly and permanently modify personal information, in accordance with generally accepted best practices, to ensure that no individual can be identified from the information, whether directly or indirectly, by any means. (Section 2(1))
- Bill C-27 contains additional provisions relating to the process of de-identifying and anonymizing personal information:
  - **CPPA** - An organization that de-identifies personal information must ensure that any technical and administrative measures applied to the information are proportionate to the purpose for which the information is de-identified and the sensitivity of the personal information. (s. 74)
  - **AIDA** - A person who carries out any regulated activity and who processes or makes available for use anonymized data in the course of that activity must, in accordance with the regulations, establish measures with respect to […] the manner in which data is anonymized (AIDA, s. 6(a))

Comparison with Bill 64 and PHIPA

- Comparison to Quebec's *An act respecting the protection of personal information in the private sector*, as amended by Bill 64:
  - "information concerning a natural person is anonymized if it is, at all times, reasonably foreseeable in the circumstances that 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 and according to the criteria and terms determined by regulation". (s. 23)
- Compare CPPA definition of anonymized data to concept of "de-identified" under the *Personal Health Information Protection Act, 2004* ("PHIPA"):
  - "de-identify" means in relation to the personal health information of an individual, means to remove any information that identifies the individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify the individual, and "de-identification" has a corresponding meaning (PHIPA, s. 2).

Identifiability Analysis in CPPA

Khaled El Emam

21st July, 2022

Replica Analytics
AN AETION COMPANY

Direct and indirect identifiers

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)

(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics
AN AETION COMPANY

Direct and indirect identifiers

DIRECT IDENTIFIERS

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

pseudonymization

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)

(c) Copyright 2019-2022 Replica Analytics Ltd.

Direct and indirect identifiers

DIRECT IDENTIFIERS

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

INDIRECT IDENTIFIERS

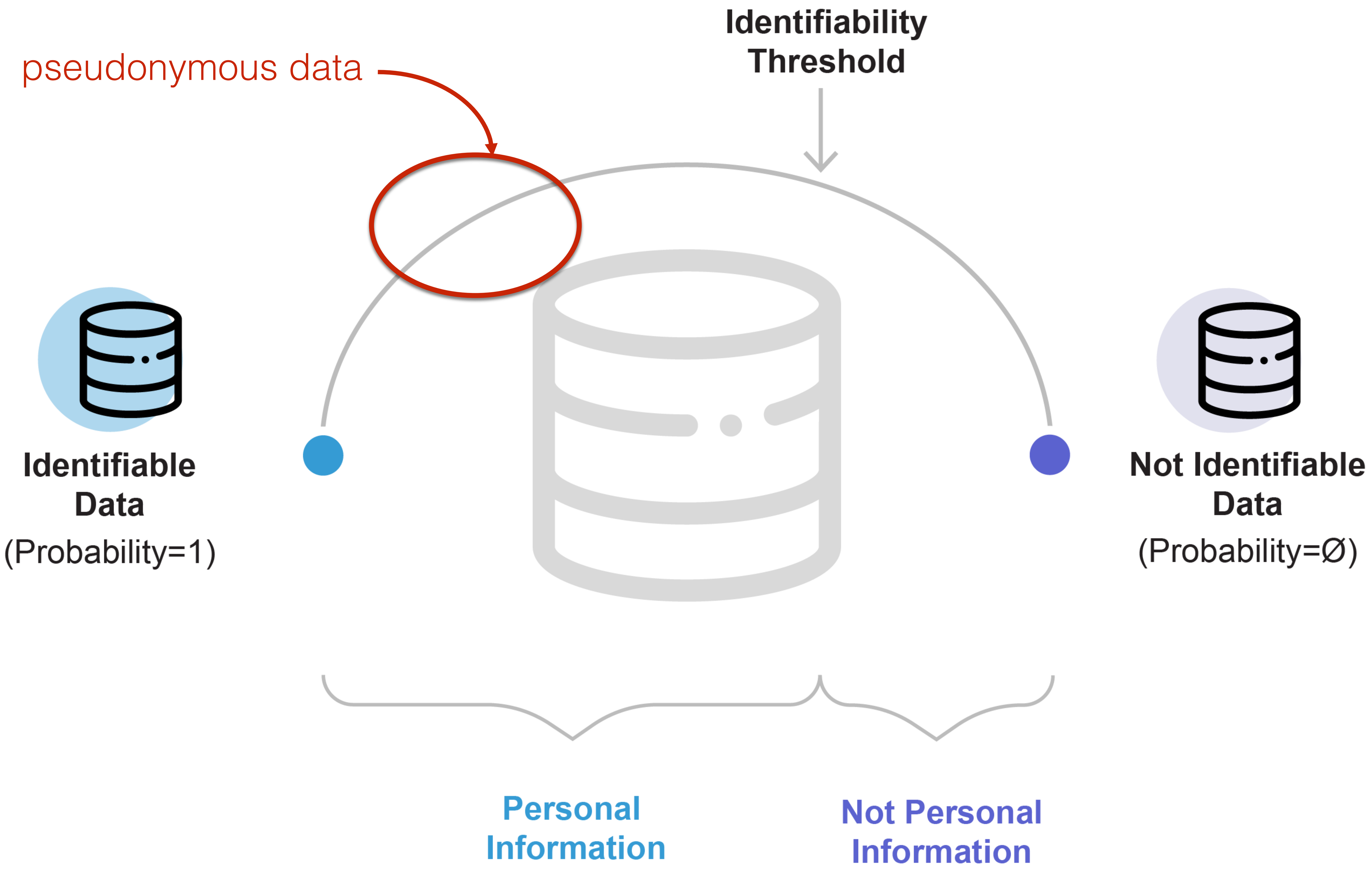
- Postal code / ZIP code
- Age / DoB
- Race / ethnicity / language
- Income
- Visible characteristics (e.g., mobility devices)
- Dates of important events (e.g., marriage, death)

anonymization

(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics
AN AETION COMPANY

Identifiability spectrum and risk thresholds



(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics
AN AETION COMPANY

The CPPA definition of de-identify can be interpreted as pseudonymize

de-identify means to modify personal information so that an individual cannot be directly identified from it, though a risk of the individual being identified remains.

==

pseudonymize

(c) Copyright 2019-2022 Replica Analytics Ltd.

The CPPA definition of anonymize sets a high standard

anonymize means to irreversibly and permanently modify personal information, in accordance with generally accepted best practices, to ensure that no individual can be identified from the information, whether directly or indirectly, by any means. […]

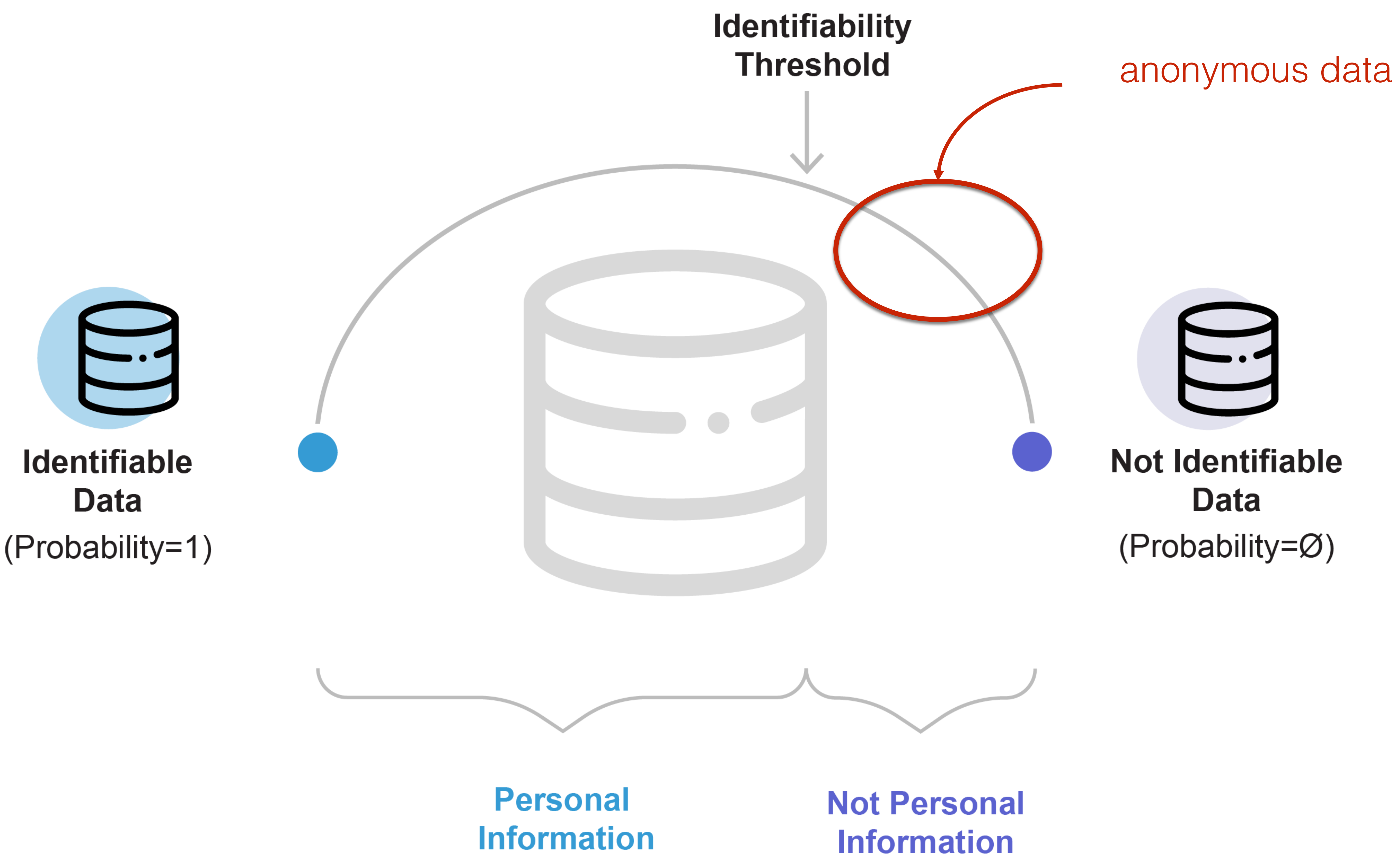
For greater certainty, this Act does not apply in respect of personal information that has been anonymized.

(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics

AN AETION COMPANY

Identifiability spectrum and risk thresholds



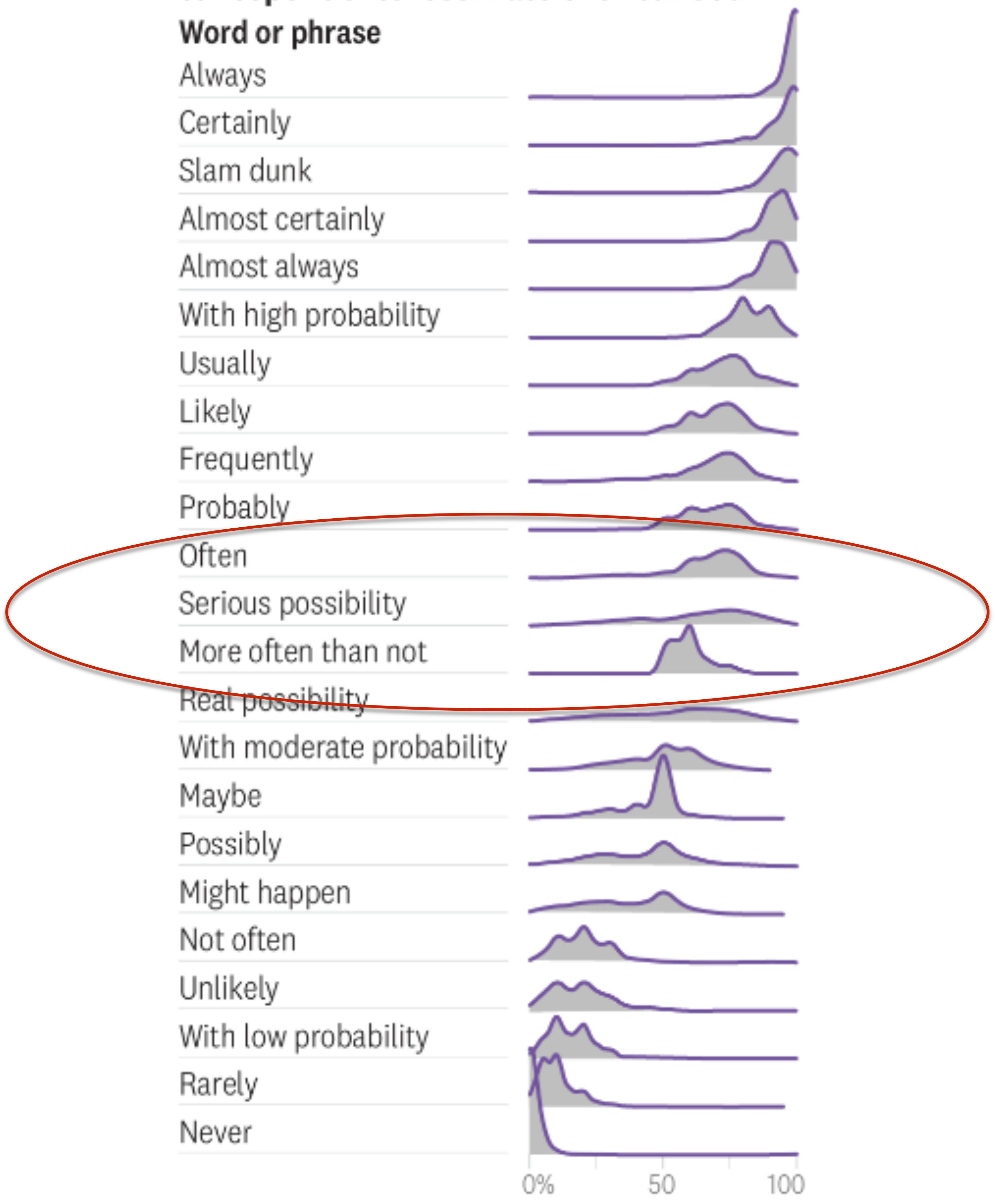
(c) Copyright 2019-2022 Replica Analytics Ltd.

How does this relate to “serious possibility” ?

How People Interpret Probabilistic Words

“Always” doesn’t always mean always.

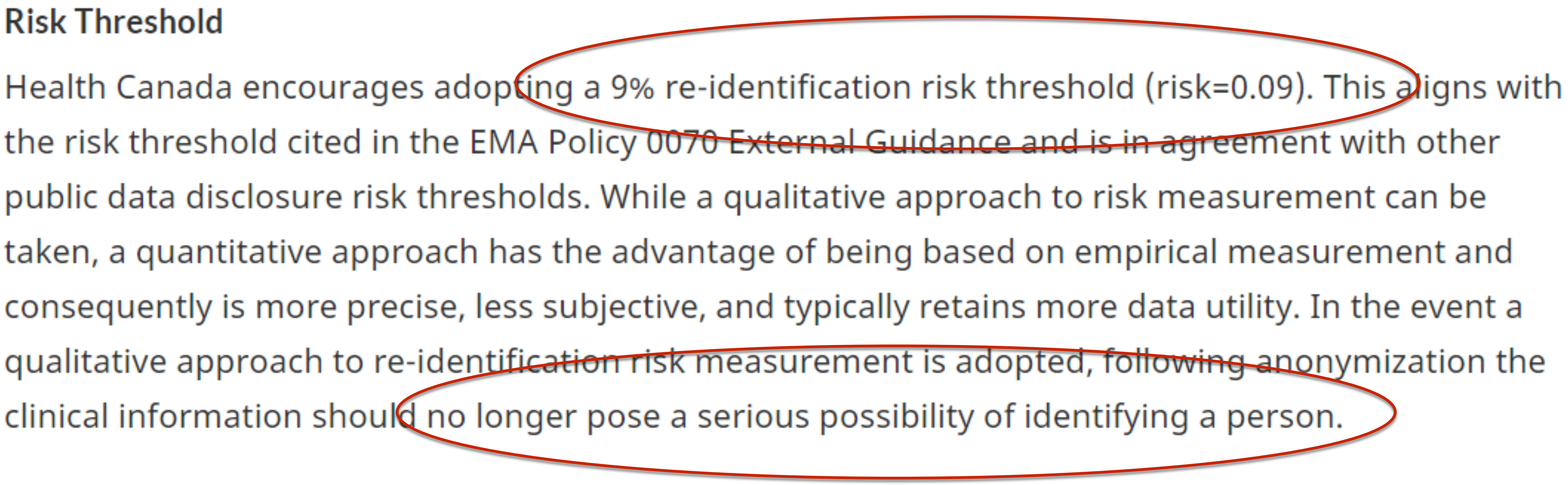
Distribution of responses according to respondents’ estimate of likelihood



Source: Andrew Mauboussin and Michael J. Mauboussin HBR

(c) Copyright 2019-2022 Replica Analytics Ltd.

How does this relate to “serious possibility” ?



Risk Threshold

Health Canada encourages adopting a 9% re-identification risk threshold (risk=0.09). This aligns with the risk threshold cited in the EMA Policy 0070 External Guidance and is in agreement with other public data disclosure risk thresholds. While a qualitative approach to risk measurement can be taken, a quantitative approach has the advantage of being based on empirical measurement and consequently is more precise, less subjective, and typically retains more data utility. In the event a qualitative approach to re-identification risk measurement is adopted, following anonymization the clinical information should no longer pose a serious possibility of identifying a person.

(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics

AN AETION COMPANY

Some considerations about risk

- Setting a threshold of zero means that all data is personal data as that is not a practical standard
- Considering any third party as a potential adversary (as opposed to anticipated data recipient) raises some questions:
    - Is this a third party who can plausibly get the anonymized data or any third party even if they cannot get the anonymized data ?
    - Does a third party who would not get the data but can follow a lawful process to get the data count (e.g., law enforcement) ?
    - Do nation states count ?
- Interpreting any means to de-anonymize:
    - Are criminal means considered ?
    - If identifiable data exists somewhere, does that imply that a means exist to de-anonymize in theory ?

(c) Copyright 2019-2022 Replica Analytics Ltd.

Replica Analytics

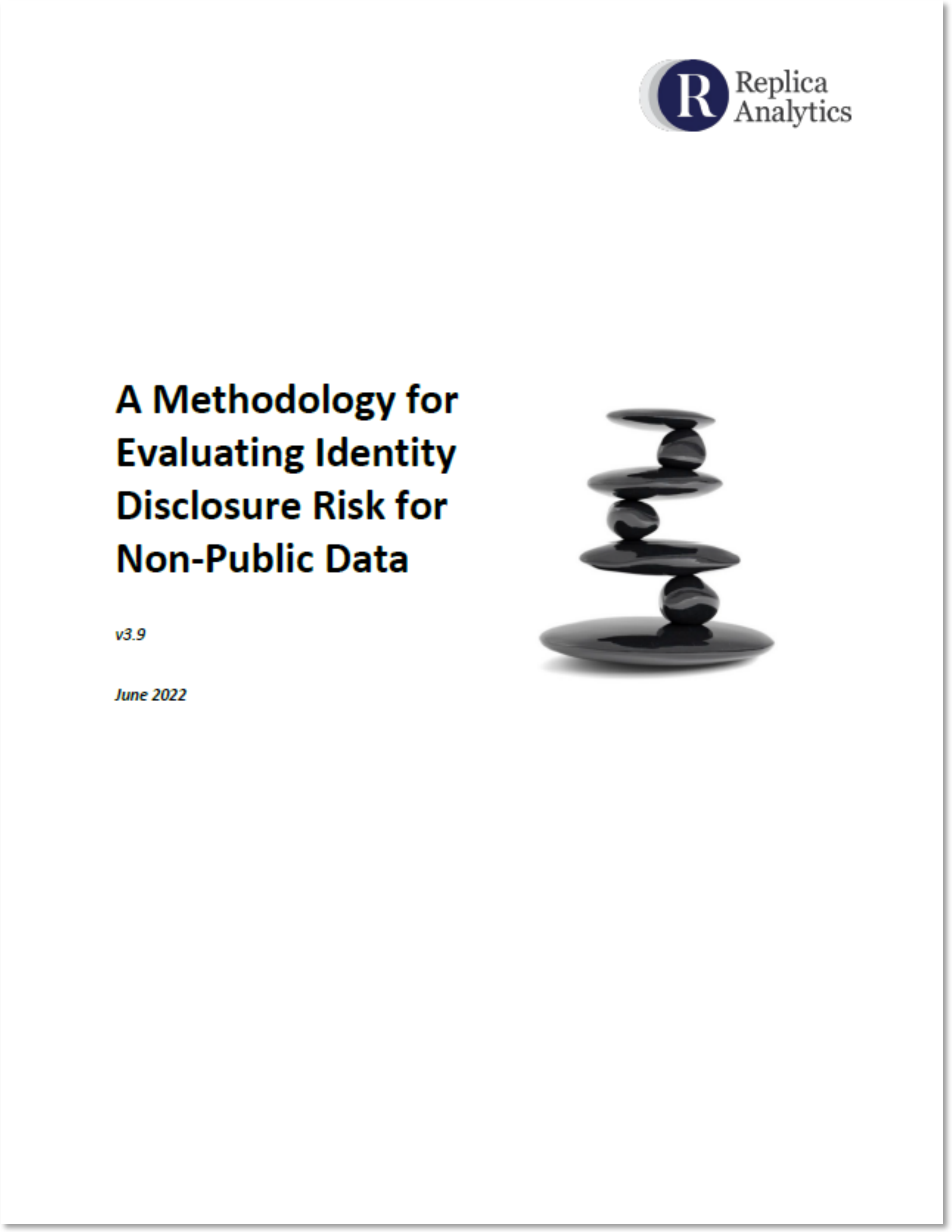
AN AETION COMPANY

Practical interpretation of “generally accepted best practices”

- Use precedents to define a threshold
- If a third party is unlikely to get the anonymized data then they would generally not be considered as an adversary
- Inform risk assessments using known re-identification attacks (i.e., not what can theoretically happen but based more empirically with a contingency / buffer)
- Law enforcement and nation states as potential adversaries are more complicated
- If the organization holds identifiable data then appropriate separations between the organizational unit that processes identifiable data and the organizational unit that processes anonymized data must be established, documented, and enforced
- Assume that contractual controls generally work
- Generally, do not make public statements about processing anonymized data or anonymization
- Controls should be used to manage *residual* risks and not be the main approach for managing risks
- Follow a fully documented process for risk management

(c) Copyright 2019-2022 Replica Analytics Ltd.

Our methodology for identity disclosure risk assessments



A Methodology for Evaluating Identity Disclosure Risk for Non-Public Data

v3.9

June 2022

- Establishes a detailed responsible process for assessing and managing identity disclosure risk

(c) Copyright 2019-2022 Replica Analytics Ltd.

Additional risks that may be relevant depending on the privacy enhancing technology that is being used

- Identity disclosure
- Attribution disclosure
- Membership disclosure

(c) Copyright 2019-2022 Replica Analytics Ltd.

Limits on disclosure of de-identified information

The CPPA authorizes an organization to disclose de-identified data to the four categories of entities identified below if the disclosure is made for socially-beneficial purposes. The four categories of recipients are:

1. a government institution or part of a government institution in Canada,
2. a health care institution, post-secondary educational institution or public library in Canada,
3. any organization that is mandated, under a federal or provincial law or by contract with a government institution or part of a government institution in Canada, to carry out a socially beneficial purpose,
4. any other prescribed entity

“Socially beneficial purposes” are defined in Subsection 39(2) as meaning:

> A purpose related to health, the provision or improvement of public amenities or infrastructure, the protection of the environment or any other prescribed purpose.

(c) Copyright 2019-2022 Replica Analytics Ltd.

However …

- Many public sector privacy laws are very old and need to be updated with respect to their handling of personal information.
- Should organizations that disclose de-identified personal information for socially-beneficial purposes as set out above enter into a contract with the recipient requiring that the recipient maintain certain privacy controls over the de-identified data (e.g., prohibition against re-identification; limiting the use to the socially-beneficial purpose(s) for which it was disclosed) ?

(c) Copyright 2019-2022 Replica Analytics Ltd.

QUESTIONS

cc: an untrained eye - https://www.flickr.com/photos/26312642@N00