CANON Canadian Anonymization Network

Data Sharing Use Cases – Quick Reference Guide

The purpose of this document is to provide a brief description of common use cases that require the application of privacy enhancing technologies (PETs). The use cases are focused on using and disclosing data for secondary purposes. No specific assumptions are made about these secondary purposes (e.g., for the public good, or to market a product).

The use cases are intended to serve as a framework for the development of more detailed industry focused case studies. The case studies would provide a concrete context and example of how PETs are being applied, the types of protections that they provide, and the trade-offs that are being made.

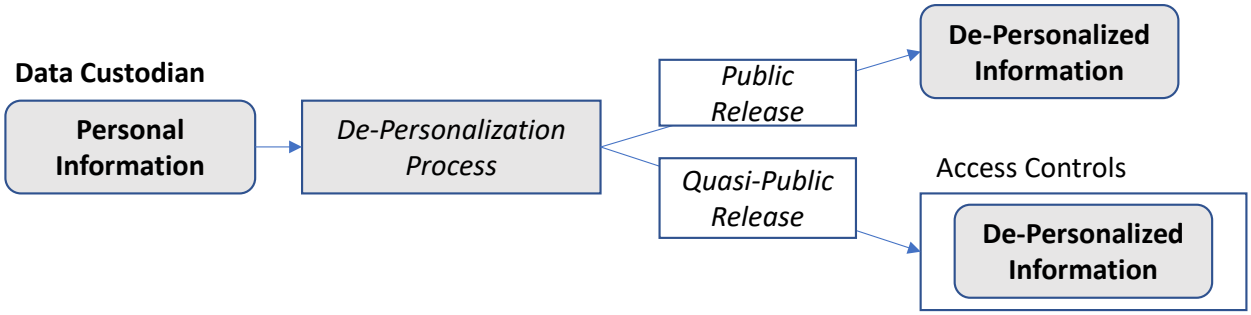
The PETs that can be used can also vary. PETs in our context are techniques for ensuring that the risk of disclosure is very small. Specific privacy metrics (such as differential privacy and k-anonymity) are used by those techniques. The PETs that we focus on as exemplars, although that is not a comprehensive list, are de-identification, data synthesis, secure computation, and federated analysis. These PETs can be applied in most of the use cases, although some of the use cases are specific to a type of PET.

The use cases are also not intended to be comprehensive. They reflect common situations, but not necessarily all situations. The list of use cases may expand over time to include more use cases as they become more commonly applied in practice.

At the level of abstraction at which these use cases are described, no specific assumptions are made about the intruder. The intruder characteristics and the specific threat models will be dealt with in the context of case studies that will be produced based on this work.

This work is supported by the Office of the Privacy Commissioner of Canada's Contributions Program.

USE CASE: Open Release of De-Personalized Data



Description

After undergoing a de-personalization process, information is published, added to an open data portal, or made available via a similar mechanism.

Data may be published publicly (i.e. without restriction), or quasi-publicly, in which terms of use are established and the identity of parties accessing the data is recorded and/or verified.

Benefits and Primary Use(s)

Open release promotes data availability, which in turn can promote transparency, data equity, and innovation.

As such, open release will often be the preferred option for applications in which any residual risk to individuals is outweighed by a significant social benefit associated with data availability, such as government open data programs or the sharing of standardized artificial intelligence training sets.

Costs

As discussed in the "Risks," openly released data faces the broadest possible re-identification threat model. As such, an adequate level of de-personalization may necessarily require a significant reduction in the informational value of the data, or a significant transformation of individual data points.

Risks

Once de-personalized data is made public, the releasing organization will retain little control over it; the effectiveness of any de-personalization will be almost entirely a product of the transformation applied to the dataset. Access controls may introduce a level of deterrence, but this will depend on the perceived or actual value of a successful re-identification to the intruder.

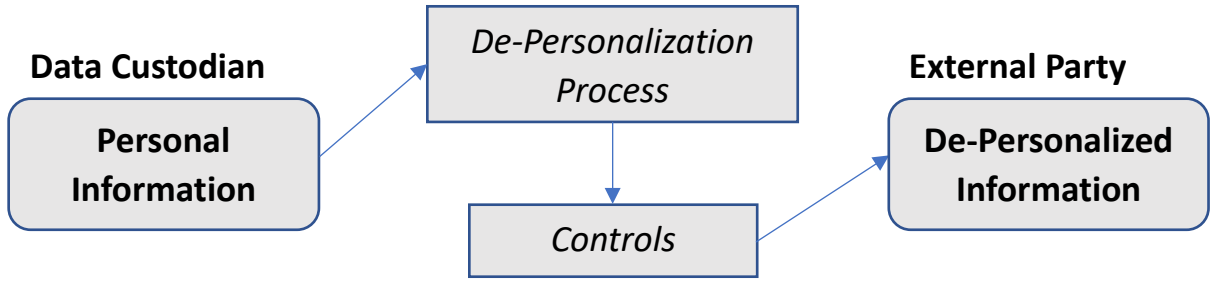
Organizations should generally assume that a re-identification attempt will occur, though the threat model adopted may vary from the moderate (the "motivated intruder," who will use publicly known techniques and public data sources to attempt re-identification, but who will have limited resources and no specialized knowledge) to the more substantial (the "malicious actor," for whom fewer restrictions apply). At least in the case of publicly released information, it should also be assumed that this attempt may occur immediately or at any future time, as parties will generally be able to maintain local copies of released data.

Organizations Should Consider:

- Releasing data in a lower level of detail – for instance, a limited number of key variables, or aggregate data – rather than record-level de-identified data.
- Continually monitoring the availability of other public datasets which may impact re-identification risk, discontinuing or modifying open release as necessary
- Releasing data in a quasi-public manner and setting conditions (such as a prohibition against re-identification attempts and/or transferring data to other parties) via clearly stated terms of use or agreements. Some effort should be expended to ensure that these terms are somewhat enforceable in practice.
- Performing a "motivated intruder test" as part of an overall risk assessment.

CANON Canadian Anonymization Network

USE CASE: Release of De-Personalized Data to an External Party



Description

The data custodian de-personalizes the information it holds, and then releases it to a single external party based on a contractual arrangement. This arrangement may include security, privacy, and contractual controls (for example, limitations on how the external party can use the data, prohibitions against re-identification attempts, and/or due diligence measures). The degree of controls is variable depending on the risk levels.

Benefits and Primary Uses

The release of de-personalized data to an external party is a very common practice – particularly where parties have an established, trusted relationship. As opposed to public release, the data custodian can establish protections through a combination of due diligence and contractual protections which can significantly lessen the likelihood of a re-identification attempt.

Costs

Thorough due diligence measures (such as auditing of data recipients) and the establishing of appropriate protections represent an upfront cost, while on-going monitoring of compliance with contractual terms can also be a significant undertaking.

Risks

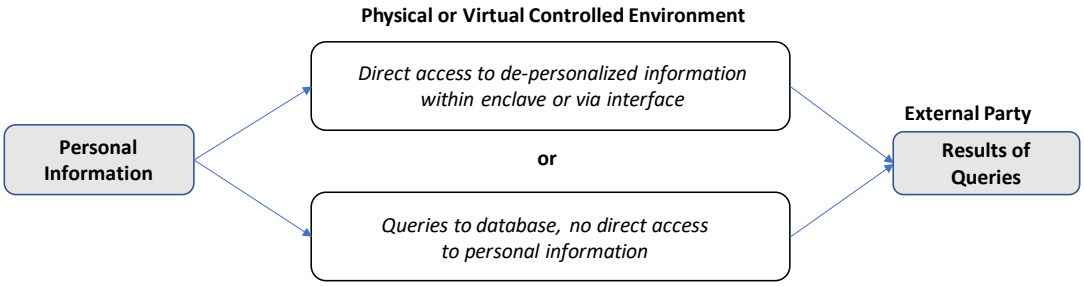
The likelihood of a re-identification attempt can be lessened by requiring privacy and security controls and contractual protections. As well, in this model once data has been released to an external party the data custodian loses control over that data; as with public release, an individual making a re-identification attempt will have the benefit of time and all reasonable external resources.

Organizations Should Consider:

- Including clear, enforceable prohibitions against re-identification in all contracts with 3rd-parties
- Perform and document due diligence on 3rd-parties, including with respect to the organization's privacy and security safeguards, their track record of handling data, any professional obligations, and so forth.
- Establishing a means of continually monitoring compliance with the terms established between the parties.

CANON Canadian Anonymization Network

USE CASE: Custodian-Controlled Access by Third-Party to Data



Description

A data custodian establishes physical or virtual space in which data can be reviewed and/or processed by external parties. External parties may be permitted to view data within this secure environment (a "secure enclave"), or only permitted access to the results of queries (a "validation server").

Particularly in virtual environments, all data access and processing will typically be monitored and recorded. Analytics results and outputs may be perturbed to manage disclosure risks. Alternatively, external parties would be permitted to only take results outside of the secure environment once they have been checked to ensure they are not disclosive.

Benefits and Primary Uses

Where a data custodian is unable – due to regulatory restriction, data sensitivity, data value, etc. – to release data (even under contractual protections), but analysis of that data would be socially or economically beneficial, a balance can be struck through controlled access.

Controlled access permits the custodian to maintain control of data (including analysis without direct access, if necessary), establish restrictions on researchers and/or research purposes, and monitoring of data processing.

Costs

Maintaining a controlled environment – including on-going decisions on what parties get access, and for what purposes – will be an on-going technical and administrative burden on the data custodian.

As well, the computing environment provided by the data custodian may be not be sufficiently powerful, may not include all desired tools, or be very costly – particularly for applications involving artificial intelligence and deep learning – creating the possibility that certain beneficial research and analysis does not take place.

Risks

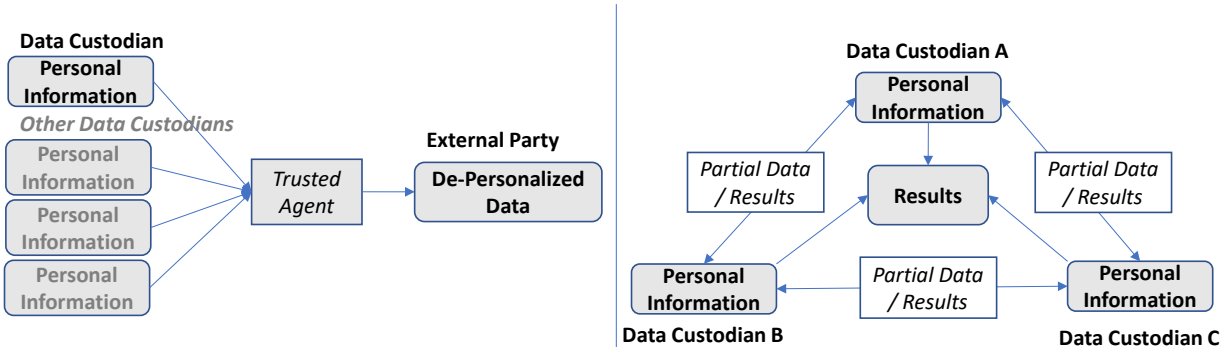
With appropriate conditions placed on parties accessing the controlled environment, and sufficient monitoring to ensure that personal information either (i) is not directly accessed and/or (ii) personal information is not permitted to leave the controlled environment, the risk of re-identification is very minimal.

However, a risk is present that important research is not undertaken if the administrative burden or use restrictions introduced by the controlled environment are unreasonably high.

Organizations Should Consider:

- Whether they have the necessary resources to support a controlled environment
- Providing an environment which supports queries being made on data, rather than raw data being provided to the external party
- Ensuring that the results of queries run by external parties are not disclosive – that is, do not reveal personal information.

USE CASE: Multi-Party Processing



Description

In a multi-party processing scenario, multiple data custodians hold personal information that is of greater value when combined; however, they are not able to simply release that data to one another.

Two potential approaches are shown above. First, the "trusted agent" approach see each of the custodians provide their information to an external trusted agent, which manages access to and use of the combined data. De-personalization can take place either before disclosure to the agent, or be done by the agent.

Second, organizations can use a technique such as "multi-party computation" which, in brief, sees each custodian perform part of a calculation on a partial set of data, and then combine their partial results into a final outcome – with each custodian having meaningful knowledge of only their own data. These are also sometimes referred to as "federated analysis".

Benefits and Primary Use(s)

The combination of data across data custodians is a powerful means of advancing economic or socially beneficial causes which could not be achieved as effectively – or at all – by a single organization. It can provide for a single point of access for researchers and others who wish to utilize data across organizations.

Costs

Establishing and maintaining a trusted agent is a significant undertaking – operational costs should not be discounted.

Similarly, the engineering costs associated with multiparty computation – including, but not limited to, establishing and validating security protocols between the parties and auditing to ensure parties remain independent – can be significant. Once deployed, changes to these protocols can be nontrivial.

Risks

When data is provided to a trusted agent by multiple parties, the potential for re-identification can increase in multiple ways. First, if not done with proper consideration, any combination of datasets will increase re-identification risk. Second, a trusted agent is likely to be the target of attack, given the value of the data they hold (as compared to any single controller).

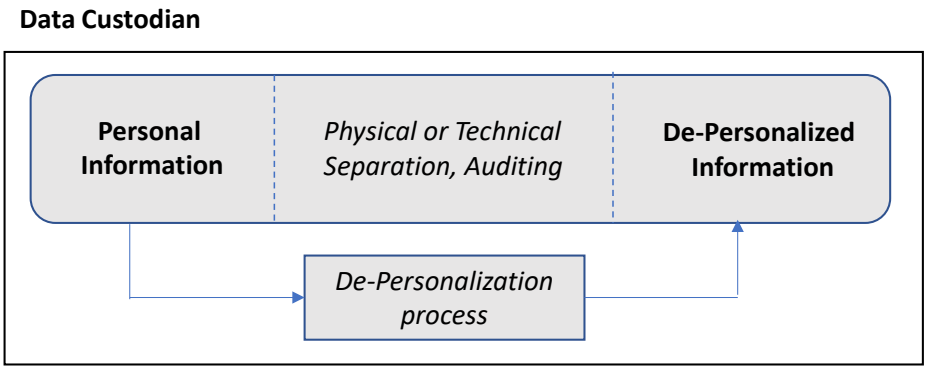
Secure multiparty computation avoids the risks posed by trusting a central agent; however, many key techniques are still under development, and the types of analytics that can be performed remain limited. Similarly, as multiparty computation requires specialized skills to develop, validate, and operate, organizations will be dependent on the still limited talent pool of human experts.

Organizations Should Consider:

Before providing data to a trusted agent, organizations should consider: what measures are in place to ensure that the risk of re-identification isn't meaningfully increased when the agent combines or releases data from multiple sources; what measures the agent has put in place to prohibit re-identification efforts by data recipients; and what measures the agent has in place to protect data provided to it.

Before entering into a multiparty computing arrangement, organizations should determine: whether the analysis they propose to undertake is amenable to this arrangement; whether the arrangement is sustainable over the long-term; and, what level of security proof is necessary for the arrangement (and, for instance, which of the "semi-honest" or "malicious actor" threat models is appropriate).

USE CASE: Internal Data Sharing



Description

Where data is used for multiple purposes within an organization, it may be the case that not each use requires access to all collected information. Thus, organizations may take the step of de-personalizing data prior to allow access to it by other departments or divisions, and establishing mechanisms (such as physical or technical separation and access logs) to prevent the recipient department from accessing the source data.

Separation measures between source and de-personalized data vary widely, and may include any of: storage in technically or physically separated databases; access limitations (e.g. an individual may only access source or de-personalized data, not both); physically separated teams; or the establishment of a subsidiary organization to hold de-personalized data.

Benefits and Primary Use(s)

Applying de-personalization to internal data can allow it to be used for internal analytics, testing, product development, etc., in a privacy-preserving manner. It is also a good demonstration of an organization's commitment to key privacy principles which underpin most legislation, such as data minimization and least-access.

Costs

The costs of this approach tend to be administrative, but within the realm of "good privacy practices." This includes establishing and monitoring access logs, segregating data storage within the organization, etc. These costs can increase significantly depending on the type of separation between source and de-personalized data.

Risks

When de-personalized data is shared internally, one must consider the information that may be available to an insider that is not accessible to an adversary. This may include knowledge of the specific methodology used to generate the de-personalized data, or knowledge of / access to the original data set.

Internal data sharing can also pose a degree of regulatory risk, as: (i) it is not clear what level of separation is considered appropriate by regulators, and (ii) it is not currently clear whether regulators will acknowledge a de-personalized dataset as non-personal information if the organization retains the source data, regardless of the additional protective measures in place.

Organizations Should Consider:

- Establishing strong internal protections, including firewalls between recipients of anonymized data and the original dataset; auditing access to the original dataset; and a clear, enforceable employee code of conduct which prohibits any re-identification attempts.
- Communicating with regulator(s) to understand their perspective(s) on expectations for data separation.