CANON Canadian Anonymization Network

Practices for Generating Non-identifiable Data

FINAL REPORT

FOR

OFFICE OF THE PRIVACY COMMISSIONER OF CANADA
CONTRIBUTIONS PROGRAM

March 2021

Contributors:

Khaled El Emam
Paige Moura
Vance Locton
Elizabeth Jonker
Adam Kardash
The CANON Steering Committee

Table of Contents

Executive Summary 3
1. Introduction 6
1.1. Problem Statement 6
1.2. Overview of Report 7
2. Terminology 8
3. Data Sharing Use Cases 10
3.1. USE CASE: Open Release of Non-Identifiable Data 11
3.2. USE CASE: Release of Non-identifiable Data to an External Party 13
3.3. USE CASE: Custodian-Controlled Access by Third-Party to Data 14
3.4. USE CASE: Multi-Party Processing 16
3.5. USE CASE: Internal Data Sharing 18
4. Case Studies 20
4.1. Methodology and Scope 20
4.2. Summary of Findings 21
4.3. Case Study #1 24
4.4. Case Study #2 27
4.5. Case Study #3 31
4.6. Case Study #4 36
4.7. Case Study #5 39
4.8. Case Study #6 42
4.9. Case Study #7 46
5. Design Patterns 50
6. Appendix A: Case Study Interview Questions 51
7. Appendix B: Author Bios 52
8. Appendix C: Members of the CANON Steering Committee 54
9. Acknowledgements 56
10. References 56

Executive Summary

This report is a descriptive analysis of the practices used by a sample of Canadian organizations to render their data non-identifiable. The purpose of creating non-identifiable data is to be able to use and disclose that data for secondary purposes.

The methods used consisted of on-line discussions and interviews with Steering Group and general members of CANON. The main findings covering practices, lessons, and challenges are as follows:

Practices

- **Layering of PETs:** Layering of PETS and multiple data transformation techniques, for example, the creation of synthetic or aggregate datasets from datasets already rendered non-identifiable. This multi-layering method provides an increased level of confidence for organizations by decreasing the overall risk of re-identification. This method is used to safeguard individual health and demographic information specifically.
- **Technical data suppression:** Use of technical data suppression rules that remove data points which because of their sensitivity, utility and/or outlier nature may potentially increase risk of re-identification. This technique is used by applying a series of small cell suppression rules to outlier data (e.g. data which highlights a medical reaction in less than 10 patient cases) and applies additional technical rules to ensure suppression rules cannot be undone. This method is also used to prevent certain sensitive elements of data (e.g. Social Insurance Number) from entering the organization's data lake environment.
- **Data perturbation rules:** Use of technical rules to add a desired level of noise or randomization to datasets, especially when a variable (or multiple variables) is identified as particularly unique and/or as having the ability to increase likelihood of re-identification. This method is used to 'perturb' data by rounding numbers up or down and/or adding statistical noise to a particularly sensitive dataset.
- **Risk evaluation framework:** Implementation and use of a risk evaluation framework and tiering/scoring system which guides an organization in handling data appropriate to its sensitivity level. In one example, a risk evaluation framework is utilized in order to aid in the identification of necessary levels of protection, whereby the risk evaluation framework results in the assignment of a tier level (or score) per dataset (e.g. Tier 1-4). Datasets assigned a Tier 3 rating were recommended to employ techniques to render data non-identifiable and go through the assessment again for a new tier level rating, whereas datasets assigned a Tier 4 rating would not be considered viable options for the research initiative. Datasets ranging from Tier 1 to Tier 2 were considered acceptable for research purposes and did not contain data which was re-identifiable in nature.
- **Manual re-identification stress tests**: Implementation of manual re-identification stress tests performed by highly specialized staff on incoming/outgoing datasets. These may include, for example, motivated intruder or reverse engineering tests. In one case study, as a final

safeguard quality check, the organization regularly had their data science team attempt to reverse engineer data sets prior to its release via reporting to a client. In another example, the organization regularly runs motivated intruder attacks on new (or significantly changed) aggregate data sets before they are released onto the web platform. Both approaches complement the application of PETs earlier in the process, and can arguably be seen as a form of quality control or final check on the identifiability of data.

- **Data consolidation**: Development and implementation of a data analytics policy framework, data dictionary, data catalogue, and PETs application standards (e.g., risk thresholds and rules for identifying 'toxic combinations' of data - dataset combinations that may increase the likelihood of re-identification). It is much easier to develop these when data is consolidated rather than existing in separate repositories across the organization, which can result in duplication and inconsistencies.
- **Data sharing agreements**: Implementation of written and enforceable agreements with data sharing partners which state that partners will not attempt to re-identify data. One organization requires all users of a software application which embeds the non-identifiable data to enter into an end-user agreement, outlining appropriate use and prohibiting data misuse (e.g., data selling, mischaracterization, attempting to re-identify data).
- **Security and privacy assessments:** Implementation and use of a variety of security and privacy assessments, which can be updated as needed when a project or initiative is significantly changed or new. Several participating organizations rely on security and privacy assessments in order to understand risk of re-identification and implement privacy-enhancing controls as needed.

Lessons Learned & Challenges

- **Data for good:** There is a need for increased public awareness of how data can be rendered non-identifiable so that organizations can conduct more socially beneficial data-driven initiatives, while simultaneously maintaining public trust.
- **Privacy programs:** Having a robust privacy program and governance model allows organizations to move faster and take advantage of data-driven work when opportunities arise.
- **Internal collaboration:** Working collaboratively with internal stakeholders from security, privacy and data teams is critical to executing data-driven work with commercial, customer and/or social-good benefits.
- **Risk appetite:** Defining risk appetite for an organization is not always straightforward but is required in order to achieve a balance between business mandates, data security and privacy aims and in order to make informed decisions on how to properly generate non-identifiable data.
- **Manual processes:** Rendering data non-identifiable is increasingly challenging when an automated process does not exist to do so and even more so when data

is not stored centrally. Bespoke processes requiring a high degree of human touch slow down organizations and may leave them at greater risk of incidents.

- **Privacy talent:** The talent pool for data privacy experts (e.g. privacy technologists, lawyers, policy experts, managers) is limited and this is a constant challenge for organizations.
- **Privacy law:** Privacy laws tend to treat certain data elements with a broad 'one size fits all' approach and may not always account for relevant context and controls (e.g. aggregation, motivated intruder testing) that may adequately protect data.
- **Global privacy landscape:** The global privacy regulatory landscape is not always uniform or interoperable and over time this can begin to exclude smaller organizations (who cannot afford large, complex compliance programs) from global competition.
- **Simplification:** The individuals applying PETs and rendering data non-identifiable are not necessarily experts in these methods. For example, they may be frontline workers dealing with patients. Some of the methods need to be simplified to be applicable at different stages of the workflow.
- **Control of Data:** One of the challenges for secondary data use or data sharing is control of the data and how that data will be used. The friction to benefiting from data can be driven by other concerns unrelated to privacy, but they need to also be considered when formulating a data sharing plan.

About CANON

The Canadian Anonymization Network ("CANON") is a network comprised of large-scale data custodians from across the private, public and health sectors, whose primary purpose is to promote the use of non-identifiable data as a privacy-respectful means of leveraging data for economic and socially beneficial purposes.

Co-founded by AccessPrivacy, Children's Hospital of Eastern Ontario, Symcor and Telus, CANON has quickly grown to include some of the largest data custodians from private, public and health institutions across the country.

The objectives of CANON are to:

- Share and exchange information about internationally-evolving, legal, policy and technical standards of rendering information non-identifiable.
- Develop a Canadian community of practice among stakeholders that rely on effective methods of rendering information non-identifiable for the success of their organizations across the public and private sectors.
- Educate the community at large about the effectiveness of alternative methods of rendering information non-identifiable, and meaningfully contribute to discussions about risks and opportunities.
- Identify emerging issues and challenges with rendering information non-identifiable, including re-identification risks, legal/policy constraints and ambiguities.
- Advocate for balanced legislative and policy standards for rendering information non-identifiable that enable innovative and beneficial uses of data, while reasonably protecting against foreseeable privacy risks.

1. Introduction

Data is increasingly recognized as a key driver for innovation. As stated by *Innovation, Science, and Economic Development Canada*, “Digital and data-driven technology is already empowering science, supporting innovation, and driving economic growth.”[1] The availability of data is critical to the health researcher working to prevent or cure disease, the government agency seeking to better serve its citizens, and the company creating the innovative products and services which create social value and drive the Canadian economy.

These and other benefits of data can be derived when organizations are able to make fuller use of the data in their custody and control, and even more so when data can be combined and made more widely available for use by multiple parties for economic and socially-beneficial purposes.

In order to leverage these benefits, individuals must be able to trust that their data are being used responsibly; just as the increased availability of data is critical, so too must the protection of individuals’ privacy underpin a robust digital ecosystem.

As reinforced in ISED’s “Canada’s Digital Charter in Action: A Plan by Canadians, for Canadians”: “[D]ata is a valuable resource helping to drive innovation, power machine learning, and improve services for Canadians. However, we must ensure that while we support the greater use of data we are also protecting the trust and privacy of Canadians.”[2]

The Canadian Anonymization Network (“CANON”) believes that generation of non-identifiable information provides a promising opportunity to achieve both data availability and privacy protection, allowing for a privacy-respectful means of responsibly leveraging data for economic and socially beneficial purposes.

1.1. Problem Statement

There is clear interest among regulators in the exploration of the role of non-identifiable information. This includes discussions in Innovation, Science and Economic Development (ISED) Canada’s 2019 white paper “*Strengthening Privacy for the Digital Age*”. The OPC has also identified “state of the art techniques that respect privacy, such as the use of synthetic data, differential privacy and depersonalization” as an area of particular interest for its 2020-21 Contributions Program. At the same time, the Canadian Anonymization Network (CANON) has been hearing from its members – as well as the broad array of organizations with whom CANON members interact in their day-to-day operations – a clear desire to implement strong practices for rendering data non-identifiable to help facilitate the socially- and economically-beneficial, yet privacy-protective, uses of data.

While we encourage more widespread and consistent adoption of Privacy Enhancing Technologies (PETs) that can render information to be non-identifiable, there are challenges including:

- Ambiguity in legal and policy norms and standards causing hesitancy by some organizations to render data non-identifiable and use data for innovative, economic and/or socially-beneficial purposes;
- A lack of technical guidance, best practices and/or operational resources resulting in insufficient capacity and inconsistent practices; and,

- Highly publicized re-identification incidents caused by a lack of understanding and/or implementation of best practices in rendering information non-identifiable resulting in potential reputational risks for organizations.

PETs as a field is similar to cybersecurity – strong, proven solutions exist, and these solutions are constantly being researched and improved upon. However, PETs are also only fully effective when the appropriate techniques are both known to, and correctly implemented by, the organizations that rely on them.

The ultimate objective of this project – and of CANON more generally – is to ensure that best practices for rendering data non-identifiable are both well-understood and followed within Canada, and to develop the knowledge and community of practice to enable that.

1.2. Overview of Report

In this report we describe a year-long exercise conducted by CANON to document current practices used by Canadian organizations to generate, use, and disclose non-identifiable data. These organizations represent some of the most sophisticated users of data, and have invested heavily to do so in a responsible manner. The objective of documenting and structuring these practice is to share them with a wider community, and also identify what works and what requires improvement. Some of the improvements identified will be undertaken by CANON as they fall within our mandate, but there are broader changes that are necessary and will require regulators, legislators, and data users to take a lead on.

The report starts off with a level-setting exercise in which the vocabulary, techniques and use cases for rendering information non-identifiable are defined. We then collected data from the CANON membership through a series of interviews of exemplar case studies that operationalize these use cases. The case studies cover multiple industries and demonstrate how organizations are generating, using, and disclosing non-identifiable data, as well as the benefits and challenges. We close off the report with some key design patterns which reflect the most common topologies that we saw during our interviews.

Throughout the project, the Steering Committee of CANON was involved in a series of reviews as a group and one-on-one with the project team to validate and provide feedback on the empirical findings. We also received feedback from the broader CANON membership throughout the project.

2. Terminology

At the outset we establish some common terminology that will be used throughout this report. The terms cover important concepts relevant for our use cases and case studies.

Below we define the spectrum of identifiability. A risk-based spectrum approach to defining identifiability allows for a broad range of innovative uses of information, while accounting for, and mitigating, a reasonable amount of residual risk. To that end, we define the following spectrum of identifiability with 3 specific states of information:

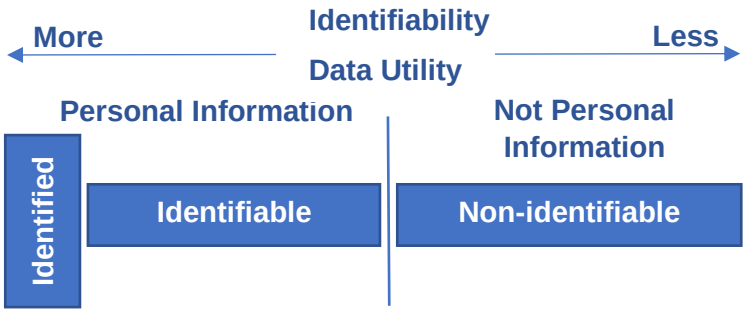


Figure 1. States of Information.

DEFINITIONS – Spectrum of Identifiability
Identified information: Information which, by itself, directly identifies an individual.
Identifiable information: Information for which there is a serious possibility in the circumstances that it could be associated with an identifiable individual.
Non-identifiable information: Information for which there is no serious possibility in the circumstances that it could be associated with an identifiable individual.

The phrase "**in the circumstances**" involves consideration of contextual factors that include: likelihood that an actor will attempt to identify the information; what other datasets are accessible to an adversary which might be combined with the information; the environment in which the information will be used or into which it will be released; and, any controls associated with the information.

Where information ultimately lies on the spectrum of identifiability will be influenced by two factors:

- the innate identifiability of the information (e.g., some information may be non-identifiable from the moment of collection); and,
- the controls applied to further reduce identifiability.

There are some additional considerations relevant to this definition:

- Rendering information non-identifiable may be achievable through multiple different techniques, including aggregation, data transformations, data synthesis, homomorphic encryption, and others. Any privacy-enhancing technology or process that creates non-identifiable information should be considered equally valid.
- Information to which the above techniques have been applied may be identifiable or non-identifiable, depending on the circumstances.
- We recommend against the use of the terms “anonymous information” and “anonymization,” because the term has been used in the past to convey a range of meanings, and in some cases to mean information for which there is no possibility of identification.

The remainder of this report will use the terms as defined here.

3. Data Sharing Use Cases

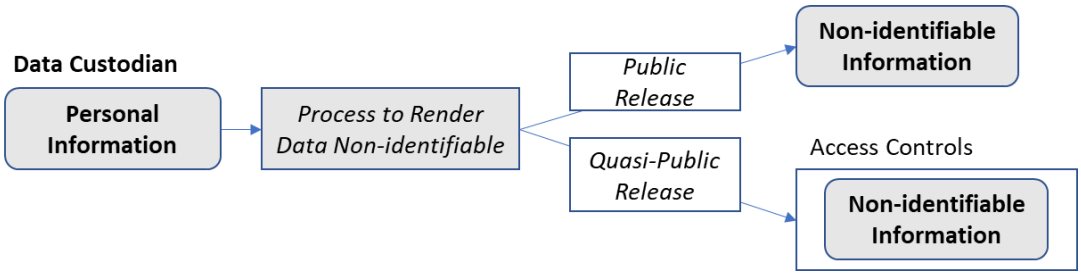
The purpose of this section is to provide a brief description of common use cases that require the application of PETs. The use cases are focused on using and disclosing data for secondary purposes. No specific assumptions are made about these secondary purposes (e.g., for the public good, or to market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

3.1. USE CASE: Open Release of Non-Identifiable Data



Description

After undergoing a process to render the information non-identifiable, 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.

Risks

Once data that has been rendered non-identifiable is made public, the releasing organization will retain little control over it; the effectiveness of any process to render the information non-identifiable 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.

Costs

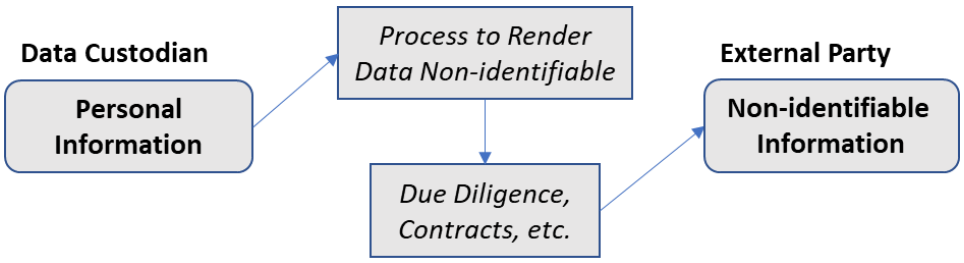
As discussed in the “Risks,” openly released data faces the broadest possible re-identification threat model. As such, to render the information non-identifiable may

necessarily require a significant reduction in the informational value of the data, or a significant transformation of individual data points.

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 data that has been rendered non-identifiable.
- 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 set 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.

3.2. USE CASE: Release of Non-identifiable Data to an External Party



Description

The data custodian renders the information it holds non-identifiable, 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 data that has been rendered non-identifiable 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.

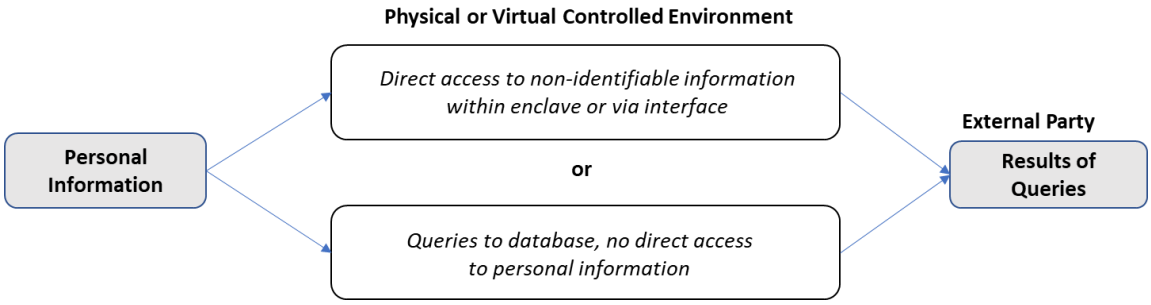
Risks

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.

3.3. 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

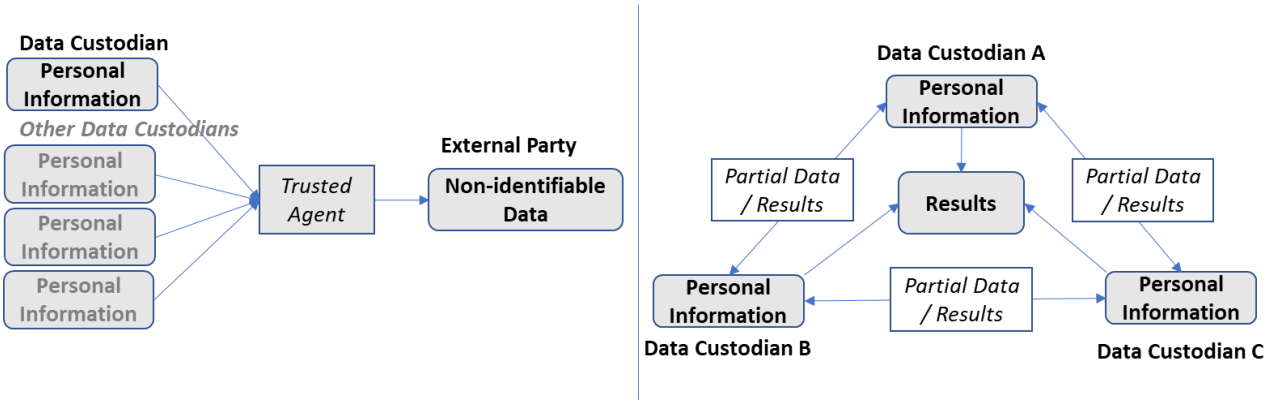
With appropriate conditions placed on parties accessing the controlled environment, and sufficient monitoring to ensure that personal information either (i) is not directly accessed and/or (ii) personal information is not permitted to leave the controlled environment, the risk of re-identification is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

3.4. 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. The information can either be rendered non-identifiable before disclosure to the agent, or this can 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).

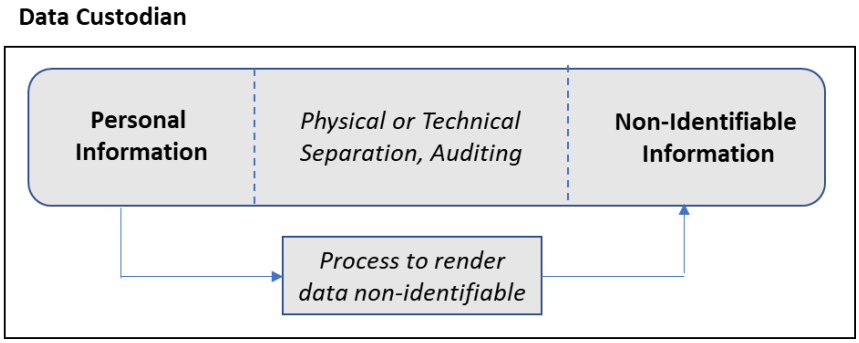
Secure multiparty computation avoids the risks posed by trusting a central agent; however, many key techniques are still under development, and the types of analytics that can be performed remain limited. Similarly, as multiparty computation requires specialized skills to develop, validate, and operate,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

3.5. 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 rendering the information non-identifiable prior to allowing 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 non-identifiable 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 non-identifiable data, not both); physically separated teams; or the establishment of a subsidiary organization to hold the data that has been rendered non-identifiable.

Benefits and Primary Use(s)

Rendering internal data non-identifiable 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 non-identifiable data.

Risks

When data that has been rendered non-identifiable 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 render the data non-identifiable, 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 dataset that has been rendered non-identifiable 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 data that has been rendered non-identifiable 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.

4. Case Studies

Based on the use cases described in the previous section, we developed a series of case studies to illustrate how they can be operationalized in practice.

Each case study reflects a real-life example of how data sharing use cases can be applied to permit privacy-enhanced data sharing and use. The following table provides an overview of how the case studies were mapped to data sharing use cases and participant industries.

Case Study - Title	Case Study - Participant Industry	Data Sharing Use Case - Title
Case Study #1 - Digital solutions for the spread prevention and treatment of COVID-19.	Healthcare	Custodian-Controlled Access by Third-Party to Data
Case Study #2 - Key Performance Indicator (KPI) design and data transformation tactics for securely reporting non-identifying data on patient support programs to pharmaceutical companies.	Technology	Release of Non-Identifiable Data to an External Party
Case Study #3 - 360-view of clients for value delivery.	Financial Services	Internal Data Sharing
Case Study #4 - Use of data-driven insights to inform credit provisioning strategies.	Financial Services	Multi-Party Processing
Case Study #5 - Utilizing aggregate data for the optimization of patient care and management.	Healthcare	Custodian-Controlled Access by Third-Party to Data
Case Study #6 - Use of analytics to understand the impacts of COVID-19 on trade activity and commercial transportation economic recovery.	Transportation & Logistics	Open Release of Non-Identifiable Data
Case Study #7 - Predicting customer perceptions of telecommunications service reliability.	Telecommunications	Multi-Party Processing

Table 1: Mapping of Case Studies to Data Sharing Use Cases.

4.1. Methodology and Scope

Given the inherent nature of the problem statement at hand, CANON selected the case study as the qualitative inquiry approach [3] which would most effectively inform the end analysis and ultimate project objective.

Interview participants from the CANON Steering Committee and the broader CANON membership were selected such that a range of professional industries were included, and a sufficient coverage of data sharing use cases was achieved. Participants were provided an interview pre-read document and interviews were conducted using a standard interview guide to support the summary analysis of findings. See *Appendix A - Case Study Interview Questions* for a full listing of the questions used during interviews. Interview recordings were transcribed, and an information extraction process undertaken, in order to identify and categorize themes per case study.

Each case study detailed within provides the reader with the following:

- The rationale and intended benefits (including the socially or economically beneficial purposes being pursued);
- Why is it important that information that has been rendered non-identifiable be allowed to be shared and/or used and how this was made possible using privacy-enhancing technologies;
- What challenges (technical, operational, legal and/or policy) were encountered (if any); and
- What lessons were learned and can be reused in similar scenarios.

4.2. Summary of Findings

Based on the case studies presented herein, the following provides a summarized understanding of commonly identified technical, administrative and/or operational **methods for ensuring data is made, and remains, non-identifiable:**

- **Layering of PETs:** Layering of PETS and multiple data transformation techniques, for example, the creation of synthetic or aggregate datasets from datasets already rendered non-identifiable. This multi-layering method provides an increased level of confidence for organizations by decreasing the overall risk of re-identification. In Case Study #1, this method is used to safeguard individual health and demographic information specifically.
- **Technical data suppression:** Use of technical data suppression rules that remove data points which because of their sensitivity, utility and/or outlier nature may potentially increase risk of re-identification. In Case Study #2, this technique is used by applying a series of small cell suppression rules to outlier data (e.g. data which highlights a medical reaction in less than 10 patient cases) and applies additional technical rules to ensure suppression rules cannot be undone. In Case Study #3, this method is used to prevent certain sensitive elements of data (e.g. Social Insurance Number) from entering the organization's data lake environment.
- **Data perturbation rules:** Use of technical rules to add a desired level of noise or randomization to datasets, especially when a variable (or multiple variables) is identified as particularly unique and/or as having the ability to increase likelihood of re-identification. In Case Study #4, this method is used to 'perturb' data by rounding numbers up or down and/or adding statistical noise to a particularly sensitive dataset.
- **Risk evaluation framework:** Implementation and use of a risk evaluation framework and tiering/scoring system which guides an organization in handling data appropriate to its sensitivity level. In Case Study #7, a risk evaluation framework is utilized in order to aid in the identification of necessary levels of protection, and includes a workflow and key questions, such as:
    - Are inputs to the dataset made publicly available?
    - Does the data identify living individuals?

    - Will you be generating new Personal Information (PI) about individuals?
    - Would disclosure represent a significant threat to personal safety, health or security of the data subjects?
    - Is it possible to re-identify individuals from the dataset?

  In this example, the risk evaluation framework results in the assignment of a tier level (or score) per dataset (e.g. Tier 1-4). Datasets assigned a Tier 3 rating were recommended to employ techniques to render data non-identifiable and go through the assessment again for a new tier level rating, whereas datasets assigned a Tier 4 rating would not be considered viable options for the research initiative. Datasets ranging from Tier 1 to Tier 2 were considered acceptable for research purposes and did not contain data which was re-identifiable in nature.
- **Manual re-identification stress tests**: Implementation of manual re-identification stress tests performed by highly specialized staff on incoming/outgoing datasets. These may include, for example, motivated intruder or reverse engineering tests. In Case Study #4, as a final safeguard quality check, the organization regularly had their data science team attempt to reverse engineer data sets prior to its release via reporting to a client. In Case Study #6, the organization regularly runs motivated intruder attacks on new (or significantly changed) aggregate data sets before they are released onto the web platform. Both approaches complement the application of PETs earlier in the process, and can arguably be seen as a form of quality control or final check on the identifiability of data.
- **Data consolidation**: Development and implementation of a data analytics policy framework, data dictionary, data catalogue, and PETs application standards (e.g., risk thresholds and rules for identifying 'toxic combinations' of data - dataset combinations that may increase the likelihood of re-identification). It is much easier to develop these when data is consolidated rather than existing in separate repositories across the organization, which can result in duplication and inconsistencies. This point was illustrated in Case Study #3.
- **Data sharing agreements**: Implementation of written and enforceable agreements with data sharing partners which state that partners will not attempt to re-identify data. In Case Study #5, the organization requires all app users to enter into an end-user agreement, outlining appropriate use and prohibiting data misuse (e.g., data selling, mischaracterization, attempting to re-identify data).
- **Security and privacy assessments:** Implementation and use of a variety of security and privacy assessments, which can be updated as needed when a project or initiative is significantly changed or new. Several participating organizations rely on security and privacy assessments in order to understand risk of re-identification and implement privacy-enhancing controls as needed.

Similarly, based on case studies collected, the following provides a summarized understanding of **challenge themes and lessons learned** for participating organizations as they looked to execute on case study objectives involving the use of non-identifiable data:

- **Data for good:** There is a need for increased public awareness of how data can be rendered non-identifiable so that organizations can conduct more socially beneficial data-driven initiatives, while simultaneously maintaining public trust.
- **Privacy programs:** Having a robust privacy program and governance model allows organizations to move faster and take advantage of data-driven work when opportunities arise.
- **Internal collaboration:** Working collaboratively with internal stakeholders from security, privacy and data teams is critical to executing data-driven work with commercial, customer and/or social-good benefits.
- **Risk appetite:** Defining risk appetite for an organization is not always straightforward but is required in order to achieve a balance between business mandates, data security and privacy aims and in order to make informed decisions on how to properly generate non-identifiable data.
- **Manual processes:** Rendering data non-identifiable is increasingly challenging when an automated process does not exist to do so and even more so when data is not stored centrally. Bespoke processes requiring a high degree of human touch slow down organizations and may leave them at greater risk of incidents.
- **Privacy talent:** The talent pool for data privacy experts (e.g. privacy technologists, lawyers, policy experts, managers) is limited and this is a constant challenge for organizations.
- **Privacy law:** Privacy laws tend to treat certain data elements with a broad 'one size fits all' approach and may not always account for relevant context and controls (e.g. aggregation, motivated intruder testing) that may adequately protect data.
- **Global privacy landscape:** The global privacy regulatory landscape is not always uniform or interoperable and over time this can begin to exclude smaller organizations (who cannot afford large, complex compliance programs) from global competition.
- **Simplification:** The individuals applying PETs and rendering data non-identifiable are not necessarily experts in these methods. For example, they may be frontline workers dealing with patients. Some of the methods need to be simplified to be applicable at different stages of the workflow.
- **Control of Data:** One of the challenges for secondary data use or data sharing is control of the data and how that will be used. The friction to benefiting from data can be driven by other concerns unrelated to privacy, but they need to also be considered when formulating a plan.

4.3. Case Study #1

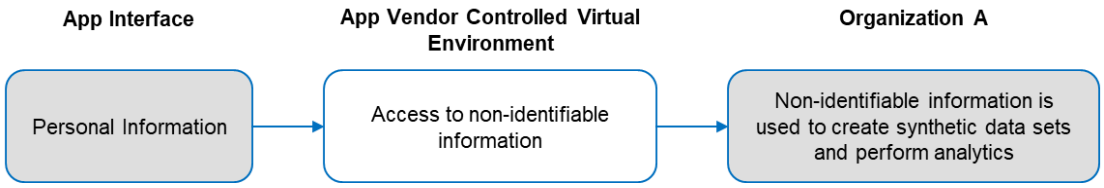
Title

Digital solutions for the prevention of the spread of and for the treatment of COVID-19.

Use Case

This case study maps to the 'Custodian-Controlled Access by Third Party to Data' use case.[1]

Data Flow



Background

In the wake of the COVID-19 global pandemic, the participating organization (referred to here as 'Organization A') worked closely with an application ('app') vendor in order to collect the Personal Information (PI) and Personal Health Information (PHI) of individuals in Canada who either, (a) suspected they had contracted COVID-19, or (b) were confirmed to have contracted the virus. PI/PHI was made non-identifiable by the app vendor so that analysis performed by Organization A and their partners had a very small risk of privacy-related exposures. Due to the sensitive nature of the data collected, extra data protection precautions (e.g. the creation of non-identifiable datasets, followed by the later creation of synthetic datasets) were put in place.

Rationale

Critical needs existed to help stop the spread of COVID-19, optimize healthcare resources (e.g. time, supplies) and to collect data that can be used by analysts and researchers globally to apply learnings at a rapid pace.

Purpose

The purpose of this work was to help prevent the spread of COVID-19 (e.g. via in-app education of the general public), while also looking for ways to improve treatment plans and save the healthcare industry and patients both time and money (e.g. via the minimization of 'trial-and-error' approaches to treatments, especially for populations with preexisting conditions or specialized needs).

[1] For this case study, Organization A represents the 'Third Party', while the application vendor represents the 'Custodian'.

Potential Benefits

Several potential benefits were identified as part of this work, including:

- Augmentation of screening practices at COVID-19 assessment centres for better patient experience and treatment plans (e.g. personalized treatment plans).
- Identification of populations requiring more personalized and/or specific products and/or services.
- Development of a sense of individual empowerment to receive better, and faster healthcare treatment plans for themselves and/or loved ones.
- Creation of a knowledge pool on the novel virus, which (over time) presents efficiencies in time/cost to the healthcare system and its patients (e.g. capacity management).
- Development of confidence (over time) among the general public that data sharing initiatives have *tangible* benefits and that organizations can be trusted to conduct data-driven work while protecting individual privacy.

Method

For the purposes of the data-driven initiative, an app was provided to individuals who suspected they may have contracted (or who had contracted) COVID-19. The use of this app by individuals was entirely voluntary.

In terms of elements of data, medical history (e.g. medications taken on an ongoing basis) and demographic information (e.g. age, race) were key, as well as data that would help researchers understand if the virus had a cyclical nature (e.g. feeling better followed by a period of feeling worse).

The app acted as a record by which individuals could keep track of symptom characterization over time and was built to collect the key pieces of data that COVID-19 assessment centers typically need. In developing the app, Organization A spoke to internal and external experts to understand what exact questions would help best characterize peoples' pre-existing health and ongoing symptoms.

Individuals who utilized the app would enter symptom-tracking data and proactively consent to the donation of this data (in aggregated form) to an open science domain, where population-specific solutions could begin to be identified. Note that if individuals did *not* want to donate their data for this purpose, they had the option of still utilizing the app for its symptom tracking functions. In support of local clinical trials, the app also featured links to ongoing trials so that individuals (at their own discretion) could participate and contribute to ongoing scientific learnings about the virus.

Users were made aware of the app through patient foundations within Organization A's network, in addition to media outlets (i.e. news stations highlighting the effort). The user experience of the app was such that individuals were notified daily when they could update the app for a new day. The app operated on the assumption it would be used regularly by individuals to track symptoms and mood.

The donated data sets were made available to machine learning institutes and other advanced analytics organizations as part of Organization A's data science coalition. As a result of this sharing, hundreds of solutions were developed and shared in the form of algorithms, infographics and statistical insights. The non-identifiable data sets were also

deployed to the general public and government administrators across the globe in order to inform social distancing practices (among other spread-of-virus prevention tactics).

Privacy Risks and Intended Safeguards

Data collected with consent via the app was made non-identifiable by the third-party app vendor. In addition, to further protect data of a sensitive nature synthetic datasets were created from the already non-identifiable datasets. Typically, organizations will use one method or the other (non-identifiable or synthetic datasets), however, Organization A utilized the 'doubly-safe' method due to the sensitivity level of the data.

Organization A worked proactively to embed a culture of privacy into the entire initiative. Where identity indicators could be stripped or intentionally broadened/abstracted, they were. The third-party app vendor chosen had a robust, pre-established process for rendering data non-identifiable and worked with Organization A to minimize any potential data exposures.

Security assessments were carried using Organization A's standard protocol. It is worth reiterating that Organization A was not the custodian of PI in this case study. Contracting provisions were created using standard practices for Organization A.

Discussion of Legal and Ethical Issues

With regards to legal considerations, Organization A worked with its internal subject matter experts to look carefully at Canadian, U.S. and global privacy legislations. Privacy notice and consent language was presented at the time and point of collection within the symptom tracker app and donation of data for the purposes of scientific research was entirely optional per individual app user. Organization A's website also detailed the terms and conditions of this work, and how it was practically executed.

Organization A relied on its own internal privacy framework to address data ethics considerations and found that challenges arose with regards to the time required to consider and address ethical and regulatory expectations while continuing to make progress and impact as fast as possible in a pandemic situation. Organization A also consulted with patient foundations with regards to which elements of PI that app should be set-up to collect.

In addition, Organization A found that privacy legislation today does not explicitly speak to a standard approach for rendering data non-identifiable. As a result, it becomes even more important to align with relevant business functions (e.g. security, legal, data analytics) and partners on the approach ultimately being taken.

Lessons Learned

- In the face of a lack of prescriptive privacy regulatory guidance, Organization A found that it becomes even more critical to align internally with relevant business functions, partner organizations and external privacy experts on the intended use(s) and governance of PI.
- Where socially beneficial motivators existed for the intended collection and use(s) of PI, Organization A found that they were able to move faster through their own internal checks and balances (e.g. consultation/assessment processes with groups like security and legal) than is status quo.

4.4. Case Study #2

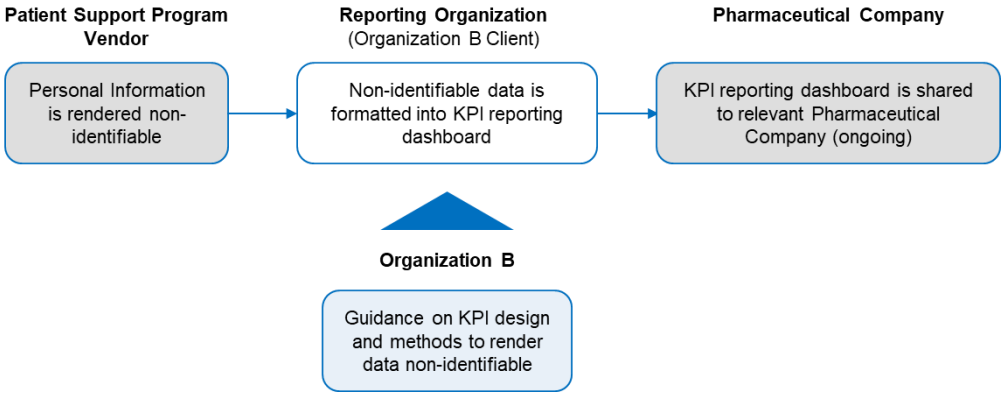
Title

Key Performance Indicator (KPI) design and data transformation tactics for securely reporting non-identifying data on patient support programs to pharmaceutical companies.

Use Case

This case study maps to the 'Release of Non-Identifiable Data to an External Party' use case.

Data Flow



Background

This advisory initiative was run on behalf of Organization B's client, referred to throughout as the 'Reporting Organization.' The Reporting Organization in this case study worked closely with vendors running global patient support programs on behalf of pharmaceutical companies. Patient support program vendors were required to regularly send KPI reports to relevant pharmaceutical companies. In order to provide this reporting, vendors worked with the Reporting Organization who rendered patient PI non-identifiable and designed a user-friendly KPI dashboard that would then be provided to pharmaceutical companies. The Reporting Organization relied on Organization B to provide best practices for KPI design and creation of non-identifiable data.

Rationale

Ongoing reporting on the effectiveness of patient support programs is needed in order to provide the best care for patients undergoing health treatments. Pharmaceutical companies require KPI-driven insights, but not necessarily PI itself, in order to fulfill their purpose of designing best-in-class patient support programs. For this reason, Organization B was engaged to advise on KPI design and tactics for rendering data non-identifiable.

Purpose

The purpose of this initiative was for Organization B to advise the Reporting Organization on how to design KPIs and implement tactics to render data non-identifiable that would protect patient privacy, while allowing for critical reporting (free of PI) to be shared back to relevant pharmaceutical companies.

Potential Benefits

The main benefit identified as part of this work was the continuous improvement of patient support program delivery, including:

- Improved patient education content and/or delivery.
- Enhanced assistance in administering medication.
- More effective patient reminders to take medication (e.g. via text messages).
- More accurate reporting regarding how patient support programs are functioning (e.g. effective, not effective).
- Insights into systemic patient complaints or issues.
- Faster remediation where improvements can be made.

Method

Data collection starts with the patient support vendors themselves who are on the ground (e.g. in-patient homes, hospitals) and interacting with patients. In order to perform their roles, they are actively collecting patient PI (e.g. health history, age, gender, race) in addition to actively monitoring their own interactions with patients. Types of interaction fields captured would include things like the number of visits made or calls taken for a patient, reasons for non-adherence to taking specific medication etc. Vendors were essentially expected to collect information regarding what worked, and what didn't work with regards to patient treatments.

Vendors, as the primary PI collectors, are trained by the Reporting Organization (informed by Organization B) on how to create and monitor KPIs. Tools like patient satisfaction questionnaires and healthcare practitioner questionnaires are utilized in order to capture key pieces of information (e.g. complaints). Information is gathered by vendors, rendered non-identifiable, translated into relevant KPIs (e.g. number of complaints reported, number of complaints by topic, % of adverse events, % of patient with complaints) and sent in summary format to the Reporting Organization. The Reporting Organization will format the KPIs into a standard KPI dashboard and share this at a regularly established cadence with pharmaceutical sponsors. From there, pharmaceutical companies can glean insights on how to continuously improve patient support programs. This entire process, from educating vendors on creating KPIs to how

PI becomes non-identifiable, is informed by the advisory role that Organization B provides for its client, the Reporting Organization.

Privacy Risks and Intended Safeguards

One of Organization B's major aims in this process was to render data non-identifiable. To this effect, all direct identifiers were removed from data sets. The non-identifiable data was then aggregated (e.g. counts, averages, percentages) in order to create KPIs. A series of small cell suppression rules were applied to ensure there are no breaches of confidentiality. As a hypothetical example, if a certain adverse reaction to a medication was present in less than 10 cases, these cases would be suppressed (removed) from a working data set in order to proactively maintain confidentiality for patients. Additional rules are applied to data sets so that suppression cannot be undone.

With regards to access, it was only certain vendor agents (e.g. nurses) that saw patient PI. Downstream, the Reporting Organization only receives non-identifiable data sets and pharmaceutical companies only see KPIs. The Reporting Organization (advised by Organization B) plays a large role in educating vendors on how to appropriately collect and apply certain rules to render PI non-identifiable. In order to proactively mitigate any risk of vendors applying data rules incorrectly, Organization B helped the Reporting Organization implement quality control checks on all datasets being sent from vendors. An analyst for the Reporting Organization was assigned to double-check all incoming data sets from vendors and will flag and send back to vendors any data sets where rules may not have been properly applied.

By design, Organization B advised that KPIs with the potential of being embarrassing or stigmatizing for patients be strictly excluded from KPI development. No socioeconomic or financial information was used to build KPI reports. The focus of KPIs was around the patient support program itself – how well it was running and how patients are responding to the program.

Discussion of Legal and Ethical Issues

With regards to legal and ethical considerations, there was a large undertaking (for which Organization B contributed) done to review vendor consent forms (on a global scale) and make them as universal as possible. This meant including plain language around what certain terms meant (e.g. aggregated) and the fact that (if patients consented) non-identifiable information would be shared back to pharmaceutical companies. Organization B helped to create a global privacy regulatory mapping to develop common terminology that the Reporting Organization could use to communicate internally and with patient support program vendors.

There was an incident during the initiative where one of the vendors started collecting data before they were instructed to do so. The data collected was necessary for the running of the program, but the vendor had not collected the proper consent in order to share back KPIs to the Reporting Organization. A consultation process was engaged to determine the best course of action. It was in the end determined that the small amount of data in question could not be utilized in KPI reporting efforts.

Lessons Learned

- The education program developed by Organization B was sufficiently useful for their direct client, the Reporting Organization, but was not appropriate for end patient support program vendors who had minimal background in the data handling processes. As a result, Organization B revisited the education program to make it more robust and with the working assumption that those doing data transformations have little to no background in data science.
- Ensure all stakeholders are onboard with critical rules and/or procedures. Just because a set of rules seems simple to one group, does not mean it will translate the same way to others.
- Re-education was required along the way because data transformation rules were not always applied correctly. Organization B found that vendors were acting in an overly conservative manner and applying rules excessively (e.g. non-identifying data to the point of being unusable).

4.5. Case Study #3

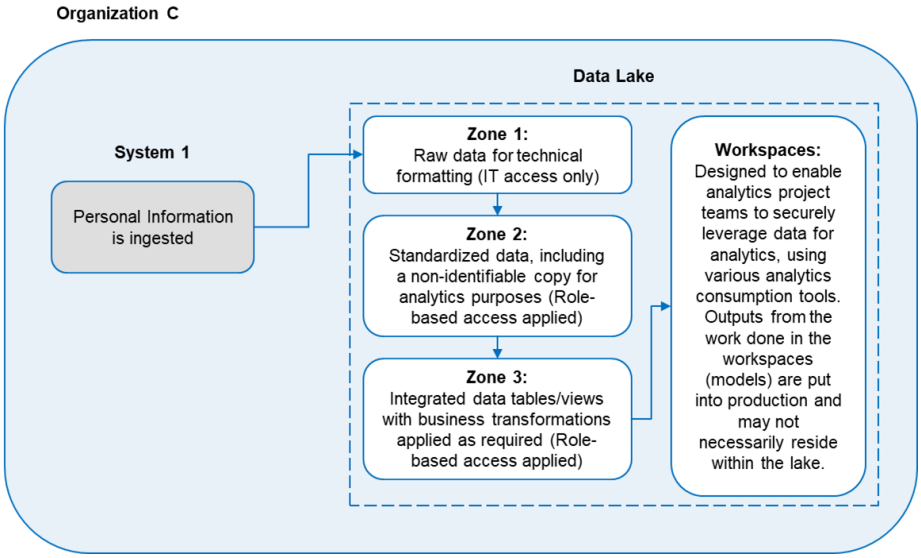
Title

360-view of clients for value delivery.

Use Case

This case study maps to the ‘Internal Data Sharing’ use case.

Data Flow



Background

In order to create a ‘360-view’ of their clients, Organization C, over the course of several years, brought down data storage silos across the organization. In the past these silos created metaphorical walls between business units who could not see across the board even if they had the same client. This at times created a poor client experience, increased the cost to service clients and left potential avenues for value creation untapped (e.g. improved client experience, retention, revenue generation). In order to have a ‘fuller picture’ of their clients, Organization C gradually moved client data into a centralized data lake. Safeguards, including the rendering of data as non-identifiable at different levels for different data lake zones, were introduced to mitigate privacy concerns and meet Organization C’s internal standards. It is worth noting that Organization C is privy to sensitive PI including, for example, information related to health benefits and retirement savings. Access provisioning as a result is strictly managed so that sensitive PI is only exposed to staff where necessary for client service fulfillment.

Rationale

As a private company, Organization C believes it has a responsibility to deliver value for its clients, including meeting its service agreements.

Purpose

The purpose of this work was to create an enhanced 360-view of clients that would enable value creation for both Organization C and its client base.

Potential Benefits

Several potential benefits were identified as part of this work, including:

- Assisting clients with achieving financial security.
- Assisting clients with living healthier lives.
- Creating opportunities for cost savings (e.g. time, money) on behalf of both Organization C and clients, for example:
    - Organization C was able to determine that certain client applicant groups existed for which specific health-related tests were not valuable because they typically came back with a consistent and predictable result. In order to gain this insight a data analysis model was created to identify these client groups (based on factors such as age, application details). This created two types of savings, one is for Organization C who avoids the cost of testing and additional analysis of testing results. The other benefit or savings is for clients who experience the increased convenience of not going through unneeded testing to get to a service/product agreement with Organization C.
    - Organization C was also able to determine that certain clients were not taking advantage of their employer's matching benefits plan for retirement savings, leaving money on the table. This created three types of benefits: clients were reminded of their employer's matching plans; employers would see increased employee savings and satisfaction with benefit plans; and Organization C managed a greater pool of assets.

Method

Organization C collects client data via clients as well as external parties (e.g. medical service providers, claim case managers, advisors) which is stored in an administrative system (referred to throughout as 'System 1') as well as a call centre system. Client data is then fed through a pipeline (developed by Organization C IT) into the central data lake environment. Upon first landing, client data goes into a raw zone. The client data is then pushed to a secondary zone where it is cleansed and formatted for consistency and a process is applied to create a non-identifiable copy of this data. This process removes what Organization C calls 'direct and sensitive identifiers' (e.g. name and contact information, banking information, government issued identifiers, etc.).

Once client data sets are made non-identifiable to a level deemed sufficient for protecting client privacy for internal processing purposes, these data sets can be used to generate internal analytics. Data scientists who work with this data cannot directly re-identify clients. As Organization C's data analytics activities mature and additional data

governance controls are applied, access within zone two and three will be applied at an increasingly granular level. From the non-identifiable zone two, data scientists have the ability to pull data into centralized digital workspaces (hosted by a Cloud provider) in order to test, analyze, build models and create outputs that can be moved into production for operationalization.

Not all the work being done within the data lake will be for advanced analytics or modeling. The data lake is also used by Organization C to fulfill straightforward business reporting purposes (e.g. sales reports).

Privacy Risks and Intended Safeguards

In thinking about privacy risk management, Organization C determined that by design certain elements of client data (e.g. SIN) would be suppressed up front from inclusion in the data lake environment itself. Tailored suppression rules prevent certain elements of client data from existing in the first, raw zone. Onwards from there, an information classification schema (owned by IT) is applied in the second zone in order to strip client data sets of direct identifiers (e.g. name). Financial and health-related information will exist in the non-identifiable datasets and is cleansed of direct client identifiers. A linking key is present in the form of an internal numeric identifier in order to re-identify data as needed for operational purposes (e.g. to create communications lists). Organization C has designed these safeguards such that the chance of staff being able to re-identify a client is quite low, unless of course there is a need to process identifiable information as part of role requirements.

At the outset of establishing the central data lake environment, a security assessment was conducted on the Cloud-based environment where the data lake now resides. This security assessment is refreshed periodically. For example, if a business unit enables additional options within the Cloud this will trigger a re-opening of the original security assessment to begin a process of determining how these new options could impact overall security posture. Once this process is completed, a refreshed security assessment report is issued and calls out any key risks in a prioritized manner (e.g. high, medium, low risk). Typically, Organization C will not allow high risks to remain unmitigated. At times, Organization C will determine, involving appropriate internal parties, that they will accept certain medium or low risks based on controls in place. Organization C may also choose to revisit medium or low risks after a defined period (e.g. a year).

In order to balance business unit needs with privacy protections, not every analytics activity requires a security assessment so long as staff are working within a pre-approved environment for pre-approved purposes. In addition to security assessments, Organization C conducts quarterly monitoring of access. Where a digital project workspace has been set up within the data lake, the workspace owner (typically at the AVP or Director levels) is responsible for ensuring access is role-based and appropriate.

An additional safeguard put in place is a logging and monitoring dashboard maintained by IT. This dashboard regularly generates reports based on staff IDs which identify ‘who accessed what, when’. Currently, this is manually managed and reviewed by IT to identify any suspicious data access activities. In the future, Organization C plans to implement algorithms in order to flag suspicious activity.

Discussion of Legal and Ethical Issues

In order to enable the 360-view of clients, several internal parties (IT, Privacy, Compliance, areas of the business) worked together to determine optimal protections and how to build privacy into the design of the data lake itself. A system of 'gatekeeping' was put in place so that areas of the business could not ingest data without going through a process of standardized checks and balances. The process needed to be flexible enough to make informed exceptions and not overly limit business units, but stringent enough to mitigate unnecessary privacy risk.

With regards to legal considerations, Organization C (in consultation with Legal and Compliance areas) confirmed that using client data to improve business processes, products and services, including communicating with clients, is included in purposes stated in their publicly available privacy policy, forms, etc. As this initiative is for internal use only, Organization C determined that these efforts do not represent a 'new use' outside of their privacy policy-stated uses. As a result, Organization C did not make additional disclosures to clients or go back to clients to ask for additional consents. All advanced analytics activities will also undergo a Privacy Impact Assessment (PIA) process to review the purpose of processing, what types of data are involved and reconcile these considerations against client consents, privacy notices and internal data principles.

In looking at potential ethical dilemmas, Organization C developed several core data principles. These principles govern the handling of client data and act as guardrails for all staff members, including those who regularly work on data modeling. These principles were socialized and validated by executive leadership. Additional guidelines have been recommended internally around transparency practices, model data risk including rationale, bias, discrimination and fair treatment of clients. In its ongoing efforts to improve transparency, Organization C is exploring upcoming updates to its public-facing privacy website pages that would provide additional information around its analytics initiatives.

Lessons Learned

- Creation of the data lake environment did not create net-new risks for Organization C, but rather provided a more streamlined and centralized storage area for client data while at the same time increasing risk mitigation and maturing governance around data use. This process of centralization made applying checks and balances (e.g. access controls) easier for Organization C.
- A robust understanding of Organization C's privacy risk appetite level (determined via a lengthy internal risk review process) was required in order to achieve a balance between business mandates, data security and privacy aims. Risk can never be fully eliminated, and it's important to make informed decisions when accepting certain risks.
- It was important for Organization C to understand the needs of different business units, without customizing the data lake environment and related controls to a point that it may have become overly cumbersome to manage. Determining an appropriate balance was key.
- Collaboration among business units, IT and Privacy teams is very important, and it is key that everyone is speaking the same language, which can take time.

Communication is critical for highly collaborative initiatives such as this, and the results can be very positive.

4.6. Case Study #4

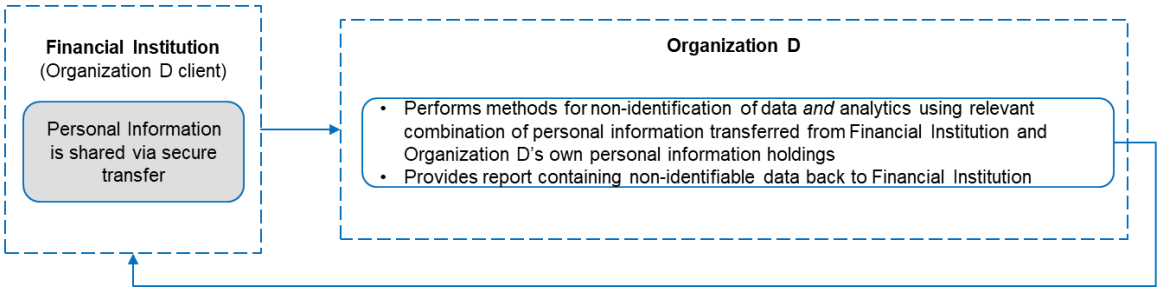
Title

Use of data-driven insights to inform credit provisioning strategies.

Use Case

This case study maps to the 'Multi-Party Processing' use case.

Data Flow



Background

Organization D regularly works with several large financial institutions (referred to throughout as 'Clients') to support their ability to analyze application data, extract insights and use these insights to inform their financial credit provisioning strategies. Based on these insights and their existing risk appetites, Clients may adjust their strategies.

Rationale

Clients want to better understand the borrowing habits of individuals in order to best assess how to distribute financial tools (e.g. lines of credit) to individuals.

Purpose

The purpose of this work was to help Clients implement the right strategies for credit provisioning and to update strategy over time, where needed.

Potential Benefits

The reinforcement of credit provisioning strategies via data-driven insights is done for several beneficial outcomes including, but not limited to:

- Minimizing manual decision-making by financial institutions.
- Preventing financial credit provisioning to individuals who (in hindsight) should have been declined upfront.
- Providing financial credit to individuals who qualify upfront in a more efficient manner.

- Avoid mistakenly declining qualified credit applicants.

Method

For Organization D to provide data-driven insights for its Clients, Clients securely share an input file containing applicant PI (e.g. name, address, phone number). Organization D may also receive additional variables (e.g. credit limit requested by the applicant, application date, product applied for) from a Client that they would like to see reflected in the end analytics report. In this scenario, the PI will not be used in the analytics activities performed by Organization D, but rather is used to go through a standard matching exercise in order to locate information about the applicants on Organization D's proprietary database.

Using the Client input data, an initial internal report is generated by a team that is organizationally separate from the data science team. This team creates the initial internal report by appending variables that pre-exist in the Organization D proprietary database, and removing the PI from the file before handing it over to data scientists. The additional Client-provided variables will be part of the file the data scientists use in their analysis.

Once the analysis has taken place, Organization D will run a process against the proposed output file. If there are quasi-identifiers in the output file, additional steps are taken to ensure the data is non-identifiable before it is returned to the Client.

Privacy Risks and Intended Safeguards

When necessary to the Client engagement, Organization D will trigger its internal process in order to protect applicant anonymity. This process was informed by an exercise in which Organization D engaged a third-party analytics consultancy to develop a framework for how to implement a more scientific approach to rendering data non-identifiable. This approach has Organization D go through several documented and repeatable steps (including a risk of re-identification assessment) to ultimately come up with an objective risk of re-identification score. This score is influence by factors such as number of input variables – the fewer input variables, the lower the likelihood of applicants being re-identified. The quasi-identifiers in the Client input file that will be returned on output also need to be examined to determine if they give rise to a likelihood of re-identification.

When a variable (or multiple variables) are identified as particularly unique and/or as having the ability to increase likelihood of re-identification, Organization D employs a variety of techniques to render datasets non-identifiable. For example, Organization D may remove the problematic variable from the output, perturb the data, bin the data or round numbers up or down. Organization D conducts this process of re-identification assessment regularly to ensure that unique examples cannot be combined to enable re-identification. As a final safeguard quality check, Organization D will regularly have their data science team attempt to reverse engineer data sets prior to its release via reporting to a Client.

As part of their business agreements with Clients, Organization D requires that Clients sign agreements which state they will not reverse engineer data to identify individuals, they will use the data for internal analytic purposes only and they will not use the information to make a decision about an individual.

Discussion of Legal and Ethical Issues

In looking at potential regulatory concerns, Organization D utilizes an internal privacy regulatory framework to ensure compliance. This framework also looks at components such as the intended use of data by Clients.

Lessons Learned

- Achieving a balance of data utility (for Clients) and preservation of anonymity (for individuals) was made possible by the implementation of techniques, such as data perturbation.

4.7. Case Study #5

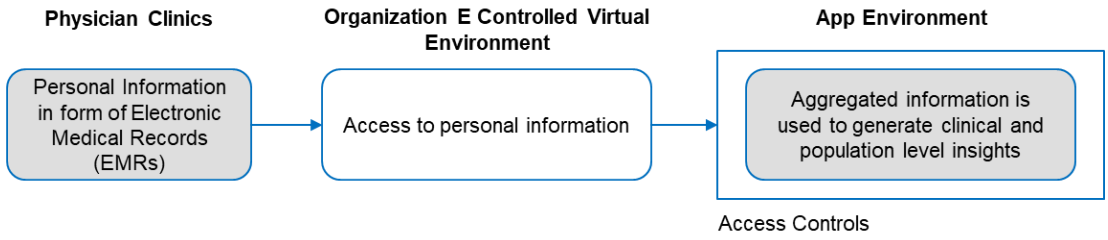
Title

Utilizing aggregate data for the optimization of patient care and management.

Use Case

This case study maps to the ‘Custodian-Controlled Access by Third-Party to Data’ use case.

Data Flow



Background

Organization E provides a service that connects physicians and other care providers with insights in the form of aggregate information for clinical management and internal quality improvement purposes. This work supports shared care, planning, and broader learnings across the health sector by enabling sharing of aggregated results across a network of care providers without compromising patient identities.

Historically, data sharing in the primary care sector has been challenging due to the sensitivity of health data, the distributed nature of data governance across the sector, legal prohibitions on sharing between the public and private organizations, and a lack of data consistency, as various Electronic Medical Record systems (EMRs) collect and present information differently. Organization E's service allows data comparisons across a range of different EMRs.

Rationale

There is a cultural shift happening in primary care, which recognizes the value of increased transparency and data sharing for public wellbeing. If physicians and clinics can learn from each other while simultaneously respecting patient privacy, there are potential benefits for all parties involved.

Purpose

To empower physicians and clinics with population-level insights to inform better patient care and management.

Potential Benefits

The benefits of this work include:

- Improved care management by enabling comparative insights that allow individual care practitioners to recognize trends and patterns in their practice (e.g. how their respective patient populations and their care management compare to community and provincial averages).
- Increased capacity for risk/benefit analysis with regards to care planning, both at the practice and health sector level.
- Improved patient outcomes at the individual and population levels.
- Enhanced potential to identify systematic challenges at different levels within the sector.
- Increased ability to leverage and share data while supporting provider autonomy and accountability for data governance.

Method

Participating medical clinics transfer patient-level EMR data to Organization E's secure, controlled virtual environment, where it is processed on behalf of the clinic. Patient-level data is aggregated to represent both the individual providers within the clinic and the clinic (all patients and providers).

The aggregated data is brought into an application environment and presented to participating clinic staff and providers through an online dashboard of clinical measurements. The intention is to give providers and clinic managers a perspective that may otherwise be challenging to see in their EMR. Providers may share and compare their aggregated results with their peers to support practice learning and planning in shared care environments.

Aggregation also occurs at population levels, when clinic-level results are aggregated a second time against all other clinics using the app. This additional summary renders the data non-identifiable, so individuals and clinics cannot be re-identified. Non-identifiable population-level data is shared with all app users to enable practical comparisons and insights at community and regional levels to support resource planning across the health sector.

Privacy Risks and Intended Safeguards

Primary care medical clinics often maintain independent EMRs, where the clinic's providers are the stewards of the EMR data. Providers have real concerns about losing control of this data (such as to the government or industry). There are fears of clinical data being used to assess performance, inform unrealistic performance benchmarks, or support agendas that put profits over patient-centered care.

Organization E's service empowers physicians and providers by assuring their continued stewardship over data through transparent service agreements and a secure architecture model. Patient-level data always remains segregated and controlled by the clinic.

Organization E requires all app users to enter into an end-user agreement, outlining appropriate use and prohibiting data misuse (e.g. data selling, mischaracterization, attempting to re-identify data). Sharing aggregate data that could identify providers or clinics is strictly controlled through meaningful and explicit consent features. Protective access provisions are designed into the app to prevent misuse of data by end-users.

Part of Organization E's strategy for rendering data non-identifiable is known as the 'rule of five.' This rule means that aggregate data groupings must contain at least five entities before becoming available through the app environment. This rule is also applied to measurement selection to protect against re-identification due to small cell sizes, outliers, or reverse-engineering attempts.

Organization E is transparent about its governance, use and sharing of the non-identifiable population-level data. It maintains a public-facing data use policy, which is agreed to by all participating clinics.

Physician oversight and a robust privacy and compliance program support Organization E in ensuring that aggregation is done at appropriate levels to create meaningful data, protect data privacy and maintain confidentiality.

Discussion of Legal and Ethical Issues

As a service provider to primary care clinics, Organization E is legally permitted to process identifiable patient data for the same purpose it was collected (e.g. to support the provision of care). Once the data is processed to support the clinic and its providers in optimizing care, Organization E relies on the clinic's permission to aggregate data to non-identifiable population levels.

Organization E relies on meaningful case-by-case consent from participating care providers to execute any sharing of aggregated measures across the network that could identify an individual provider or clinic.

Lessons Learned

- Trust and engagement with care providers are essential to leveraging primary care data for population-level analysis.
- Aggregated data can provide valuable information to support patient care, quality improvement and health system insights.

4.8. Case Study #6

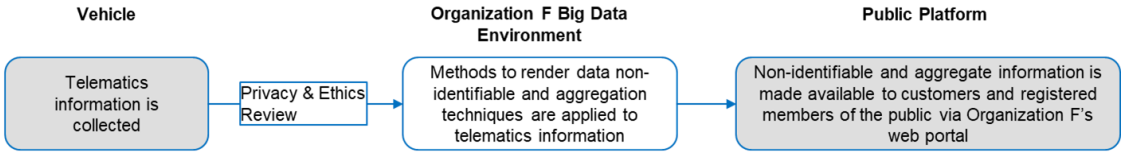
Title

Use of analytics to understand the impacts of COVID-19 on trade activity and commercial transportation economic recovery.

Use Case

This case study maps to the ‘Open Release of Non-Identifiable Data’ use case.

Data Flow



*Note: By design information collected as part of this flow is never attributed to an individual.

Background

With the onset of the COVID-19 pandemic, Organization F found it crucial for its customers, as well as members of the public (e.g. other organizations) working on economic recovery efforts, to understand the impact of the pandemic on businesses from a commercial transportation and trade activity recovery perspective.

Rationale

As economies adapt to the aftermath of COVID-19 induced lockdowns and social distancing practices, it was important for Organization F’s customers to understand the tangible impacts to commercial transportation and trade activity and where recovery is taking place.

Purpose

To understand how the global COVID-19 pandemic impacted trade activity and commercial transportation.

Potential Benefits

Several potential benefits related to data consumption and the release of non-identifiable insights were identified as part of this case study, including:

- Responding to public interest for insights.
- Providing near real-time insights to help inform public policy decision-making with regards to stimulus funding and/or other economic recovery efforts.

- Providing insights for Organization F's customers with regards to how COVID-19 may or may not have impacted them so that they could adjust for the best result.
- Benchmarking insights for Organization F's customers to understand across an industry sector if they were impacted in a similar manner (or if they are an outlier).
- Supporting improvements and efficiencies of scale.

Method

Organization F developed a business analytics solution that contained an aggregation of telematics information segmented by different industries and countries. This solution looked at aggregate information from pre-COVID, post-COVID and how countries/industries' commercial fleets began to recover. This analysis was then taken a step further to provide a view on how trade (e.g. between Canada and the U.S., the U.S. and Mexico) was impacted. By design, the information collected for this work was at no point connected to an individual (e.g. a driver).

After necessary due diligence, insights at this level are then uploaded to Organization F's web-based platform. The platform is accessible by existing customers of Organization F as well as to members of the public who go through a straightforward registration and email authentication process to gain access online. Organization F found that there were several social good related causes for which non-customers and other organizations were utilizing the resulting insights (e.g. economic recovery efforts, smart city transportation planning, road infrastructure planning, traffic management).

Privacy Risks and Intended Safeguards

As information flowed into Organization F's cloud-based environment, aggregation techniques were applied and the result was tabular data that stated the number of trips, fuel used, and border crossing broken down by jurisdiction and industry. These numbers were used to create insights in the form of percentage differences before and after COVID-19. An ethical and privacy risk review is conducted on data gathered for insights generation and external release.

To inform its regular intake and handling of data sets, Organization F implemented a data and analytics policy framework. Part of this framework addresses more technical controls and thresholds of data analysis. For example, depending on the data set, a data scientist may only go down to a country-level insight as opposed to a city-level insight. The data and analytics policies help define what those thresholds should be. In the near future, Organization F plans to roll out a new data cataloguing platform that would associate acceptable use policies with each type of data set in its custody. The platform would act as a data set lineage or 'source of truth', maintaining details such as where certain data resides and what it can be used for. Further, Organization F implemented a Privacy by Design (PbD) framework where a dedicated Privacy specialist is embedded within the Data & Analytics team to ensure Privacy preservation is considered at every step of the process.

From a privacy risk management perspective, Organization F regularly runs motivated intruder attacks on new (or significantly changed) aggregate data sets before they are released onto the web platform. This sort of testing looks at technical controls put in place to protect privacy (e.g. aggregation) and creates a scoring of sufficiency. This scoring is vetted through an approval lifecycle to ensure quality control. Where privacy-related controls are deemed non-sufficient, this information goes back to the data science team

to address. Note that data analysis has been set up so that no customer, vehicle or person is individually identified during the telematics analysis process.

In addition to technical controls, the framework also speaks to ethical considerations around *how* Organization F leverages data. Organization F operates by a 'customer first' ethos, which mandates that all data analytics should serve a purpose which is beneficial to customers regardless of whether it is deemed legally acceptable.

Organization F has a data risk assessment committee which meets to confer on a number of topics related to data and analytics, for example, which data sets to release (especially where a data use case may be in some way new or unique for Organization F to put out on its platform). Individuals from business, legal and technology sides participate in these regularly held committee meetings in order to accelerate the right actions when it comes to use of data. The most frequently employed data use cases are outlined at a high-level in Organization F's privacy policy.

Privacy training is provided to new hires and internal transfers who will work closely with data sets in their day-to-day work. New staff members are also introduced to subtle privacy considerations and are trained to think critically about the application of privacy in their work. At the end of the training there is a testing portion which staff are required to pass in order to proceed in their onboarding process. In addition to privacy training for new hires, there is also regular annual privacy training done for the entire organization. All staff are required to perform an annual attestation for privacy.

Discussion of Legal and Ethical Issues

As discussed above, Organization F conducts ethical and privacy risk reviews on datasets gathered for insights generation and external release. In addition, Organization F looks at each potential data use case with a 'customer first' as well as a 'public good' angle. Where a customer first or public good angle is not immediately apparent, Organization F would then escalate decision-making to its risk committee for a deeper conversation around cost/risk/benefit analysis.

Early on Organization F designed its internal processes and policies to address the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). This early adoption of practices (e.g. data architecture design) which support compliance with strict data privacy laws is what enabled Organization F to achieve the objectives that it did with its business analytics solution. Had they not proactively done this pre-work and fully understood legal implications regarding data residency and data use they would not have been as confident exploring the realm of possibilities with aggregated data. At times, Organization F found the broadness with which privacy law treats certain types of data challenging for their business model.

Lessons Learned

- The talent pool for data privacy experts is limited and this is a constant challenge for organizations.
- Privacy laws tend to treat certain data elements with a broad 'one size fits all' approach and do not account for relevant context and controls (e.g. aggregation, motivated intruder testing) that may adequately render data non-identifiable.

- The global privacy regulatory landscape is not always uniform or interoperable and over time this can begin to exclude smaller organizations (who cannot afford large, complex compliance programs) from global competition.
- Privacy compliance can be extremely helpful from a business perspective if it addresses a concern that customers have. Customers respond positively when you can tell them exactly what you do with their data and why.

4.9. Case Study #7

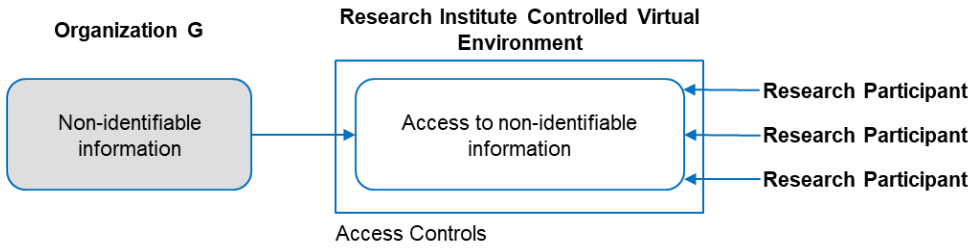
Title

Predicting customer perceptions of telecommunications service reliability.

Use Case

This case study maps to the ‘Multi-Party Processing’ use case.

Data Flow



Background

Organization G was invited to submit a challenge to a research event that brought together Artificial intelligence (AI) specialist from around the globe to work collaboratively and intensively on solving telecommunications (‘telco’) industry problems using real operator data. Organization G submitted a data study challenge around designing a solution for how to actively predict customer perceptions of telecommunications service reliability utilizing a combination of data inputs, which included:

- Key Performance Indicators (KPIs) for servicing (e.g. adverse network events, latency)
- Customer survey responses (i.e. service reliability ratings)
- Customer information (e.g. customer tenure, plan type, data usage)

Rationale

The global telecommunications space is entering an age of seeking to better understand the onboarding of, and capabilities associated with, AI. This research event represented an opportunity to do just that, and in the company of some of the worlds brightest AI experts.

Purpose

The purpose of this research initiative was to design a solution that was able to predict customer perceptions of service reliability, in order to proactively mitigate potentially negative customer perceptions, experiences and/or brand impacts.

Potential Benefits

Several potential benefits were identified as part of this work, including:

- Hands-on learnings around practical application of AI for Organization G, participating telcos and academics to take away and apply to other AI-driven work.
- A solution which can proactively monitor and assess customer perceptions of service reliability.
- The ability to monitor and address customer concerns, leading to overall higher rates of customer service satisfaction.

Method

Organization G began this initiative with designing a data study challenge and determining what types of data to involve and what problem(s) to potentially solve. Organization G then submitted their proposed data study challenge to the research institute organizing the event. Upon selection, Organization G needed to ensure the dataset that would be used during the research event was non-identifiable.

The research institute required that Organization G identify the security level (using a Tier 1-4 scale) of the proposed dataset. If, for example, a participating organization wanted to submit a data study challenge using a Tier 4 dataset the proposal would be denied on the account that the dataset represented too high a risk if used for academic purposes. To illustrate the scale range, a Tier 1 dataset would imply that this data could be made publicly available with no major privacy and/or security risks anticipated, while a Tier 4 dataset would include raw data that is plainly, or easily associated to an actual individual.

The resulting non-identifiable dataset was uploaded to a secure Cloud-based database that was made accessible to research event participants. Access to this database was limited to those participating organizations who were given access credentials via the research institute itself and only accessible from within that one physical location. The end product was a report which detailed the research effort and end solution developed.

Privacy Risks and Intended Safeguards

In planning the design of the data study challenge, Organization G went through a highly manual process of creating what was called a data dictionary. The purpose of the data dictionary was to list out column names and data types (e.g. postal code) and determine the associated appropriate method for rendering data non-identifiable. This consolidated view helped Organization G build privacy awareness into the design of its data study challenge by allowing staff to think through potential ‘toxic combinations’ of data.

Multiple methods were used to ensure data was made non-identifiable, one such technique used was to aggregate data at certain levels. For example, call time may be aggregated so that more call times fall into the same ‘bucket’ of call length, thereby making it harder to identify any one individual call – Organization G referred to this process as ‘bucketing.’ Another example was the use of a salt (random data) to hash customer phone numbers. The intention was to make it extremely difficult to re-identify any data that would be sent to the research institute for use during the event. The entire process was highly manual, demanded high levels of internal collaboration and required that a staff member inspect data and apply the agreed upon transformation techniques (e.g. bucketing).

In addition to Organization G's own internal privacy and security safeguards, the research institute hosting the event provided participants with a risk evaluation framework to aid in the identification of necessary levels of protection. The framework provided organization G with a workflow and key questions, such as:

- Are inputs to the dataset made publicly available?
- Does the data identify living individuals?
- Will you be generating new PI about individuals?
- Would disclosure represent a significant threat to personal safety, health or security of the data subjects?
- Is it possible to re-identify individuals from the dataset?

The risk evaluation framework resulted in the association of a tier level (or score) for the dataset. Datasets assigned a Tier 3 rating were recommended to employ techniques to render data non-identifiable and go through the assessment again for a new tier level rating, whereas datasets assigned a Tier 4 rating would not be considered viable options for the research initiative. Datasets ranging from Tier 1 to Tier 2 were considered acceptable for research purposes and did not contain data which was re-identifiable in nature. As part of this process, Organization G was enabled to assign the tier level of the dataset. This proposed tier level was then reviewed by an investigator from the research institute. Where discrepancies between Organization G's proposed tier level and the investigator's proposed tier level existed, a 'referee' of from the research institute team would provide an additional review and final decision. This process helped balance the interests of all parties, create productive dialogue and ultimately keeping privacy at the center of the research initiative's design.

Once Organization G had a data set which was classified as less than a Tier 4 level, they were able to use a secure transfer method to share the dataset to the research institute's controlled virtual environment.

Discussion of Legal and Ethical Issues

Organization G was proactive in providing transparency around data use to its customer base. The content made available to customers speaks to what the community benefits of this type of work are and the related performance and service benefits for customers. Prior to this research opportunity, Organization G had put in place a robust privacy program and governance model that allows for innovation while simultaneously championing customer privacy.

Lessons Learned

- There is a need for increased public awareness of how data is made non-identifiable so that organizations can do more public good with data-driven initiatives.
- Having a robust privacy program and governance model allow organizations to move faster and take advantage of data-driven work when opportunities arise.
- Working collaboratively with internal stakeholders from security, privacy and data teams is critical to executing data-driven work with commercial and customer benefits.

- Organization G faced challenges because the data feeding into its end dataset in this case was not stored centrally, nor was there a process to automate the process or rendering data non-identifiable. This required a bespoke process with a high degree of human touch to ensure data was made properly non-identifiable.

5. Design Patterns

The design patterns presented in this section reflect common topologies that appear repeatedly in our case studies. There are many possible topologies for using and disclosing non-identifiable data, and for rendering data to be non-identifiable. However, only a subset of them appear in practice. A plausible explanation is that these topologies work in that they deliver business value and also are protective of privacy. Here we distill some of these common topologies as “design patterns”.

A topology has three dimensions:

Layering of PETs (Low/High). This refers to the extent to which multiple PETs were applied to render the data to be non-identifiable.

Form of non-identifiable Data (Aggregate/Individual). This captures whether the non-identifiable information was shared in aggregate form (e.g., an analysis produced a report or KPIs) or as individual level data.

Level of Controls (Low/High). This dimension captures the level of controls (security, privacy, and contractual) that are put in place on the consumers of the non-identifiable data.

Design Pattern (DP)	Layering	Form	Controls
DP1	High	Individual	Low
DP2	Low	Aggregate	Low
DP3	Low	Individual	High

The three design patterns that emerge demonstrate three general approaches that are used.

Organizations that layer PETs are making sure that the risk of identification is very low. These organizations are willing to share individual-level data directly with external parties with fewer controls (even sharing the non-identifiable data with the public). Organizations that do not layer PETs and are sharing individual-level data directly are doing so with a high level of controls. The remaining organizations are sharing aggregate data as a report of results, for example. These do not require a high level of controls on the side of the data consumers.

These design patterns are a rational way to manage the risks. What we see is PETs layering and controls being treated as alternative ways to manage the risk when non-aggregate data is being provided. An important question then becomes what are the tradeoffs between using PETs layering versus controls? Are there economic advantages either way, for example? These are worthy questions to examine more closely in future work.

6. Appendix A: Case Study Interview Questions

- Context
    - Please confirm your organization's sector (health, private, not-for-profit, etc.).
    - With regards to the case study you've chosen to share, please describe the purpose of the data sharing, including an explanation of the role of data in fulfilling the purpose and any requisite background required to understand the case study.
- Costs / Benefits / Risks
    - At a high-level, are you able to describe the monetary or human costs of the process?
    - Additionally, can you provide an understanding of the anticipated and/or realized benefits of the process as well as any associated risks?
- Data Flows
    - At a high-level, can you provide us with an understanding of what data is involved in the process, where it originates and to where it is sent?
- Methods to Render Information Non-Identifiable
    - Please describe the methods used to render information non-identifiable
- A Priori Privacy Considerations (other than rendering information non-identifiable)
    - Other than the method(s) you've described to render information non-identifiable, were any priori privacy considerations made/put in place as part of the process (e.g. Has the organization provided notice to individuals and/or received consent for the process; has the organization used only non-sensitive information; would individuals reasonably expect that their data will be used in this manner; etc.)?
- Assumptions
    - As part of the testing process, did your organization assume a motivated intruder or a more determined adversary?
- Controls
    - After the information was rendered non-identifiable, what controls are placed on the data and/or data recipients (e.g. access controls)?
- Regulatory Challenges
    - Did your organization encounter any regulatory issues – including uncertainty about application of a legislation and/or actual interventions by regulators? If yes, how (at a high-level) were those challenges addressed?
- Lessons Learned
    - Regarding lessons learned, are there any other pieces of information you believe worth sharing?

7. Appendix B: Author Bios

Khaled El Emam

Dr. Khaled El Emam is a Senior Scientist at the Children's Hospital of Eastern Ontario Research Institute and Director of the multi-disciplinary Electronic Health Information Laboratory, conducting research on privacy enhancing technologies to enable the sharing of health data for secondary purposes, including synthetic data generation and de-identification methods. He is also a Professor in the School of Epidemiology and Public Health at the University of Ottawa.

Khaled is a co-founder and Director at Replica Analytics, a company that develops data synthesis technology. As an entrepreneur, Khaled founded or co-founded six companies involved with data management and data analytics. Prior to his academic roles, he was a Senior Research Officer at the National Research Council of Canada. He also served as the head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany.

In 2003 and 2004, he was ranked as the top systems and software engineering scholar worldwide by the Journal of Systems and Software based on his research on measurement and quality evaluation and improvement. He held the Canada Research Chair in Electronic Health Information at the University of Ottawa from 2005 to 2015. Khaled has a PhD from the Department of Electrical and Electronics Engineering, King's College, at the University of London, England.

Paige Moura

Paige Moura is a trusted privacy and data optimization consultant who has served nearly every industry. She works with her clients as they strategically navigate the data use challenges and opportunities presented by the digital economy. Paige has conducted over 50 privacy and security assessments and is consistently viewed by her clients as an effective integrator between business and data protection teams. Paige has a background in both economics and digital design, and maintains CIPM and CIPP/C designations.

Vance Locton

Vance Lockton is a Senior Technology and Policy Advisor for the Office of the Information and Privacy Commissioner of Ontario. At the time of his participation in this project, he was a Policy and Operational Consultant for the Canadian Anonymization Network. He has also worked for the Office of the Privacy Commissioner is Canada, and Waterfront Toronto (during its Quayside project with Sidewalk Labs).

Vance holds Masters degrees in computer science and public policy, and holds the CIPP/C and CIPM designations.

Elizabeth Jonker

Elizabeth Jonker is Research Coordinator and Privacy Officer of the Electronic Health Information Laboratory at the CHEO Research Institute. Elizabeth has over thirteen years of experience in privacy research, contributing to numerous research projects and co-authoring fifteen articles published in academic journals. She is a member of the IAPP and has been a Certified Information Privacy Professional (CIPP/C) since 2012.

Elizabeth has prior experience in program coordination and facilitation with the City of Ottawa. She holds an Honours BA from the Faculty of Arts, University of Ottawa from which she graduated magna cum laude.

Adam Kardash

Adam Kardash is an acknowledged Canadian legal industry leader in privacy and data management. He is chair of Osler's national Privacy and Data Management practice, and leads Osler's AccessPrivacy thought leadership platform. Adam has been lead counsel on many of the most significant privacy matters in Canada, including the largest cyber security incidents and regulatory investigations in Canada to date. He advises Fortune 500 clients in their business critical data-protection issues, compliance initiatives and data governance. Adam has extensive experience in the privacy law area and regularly advises Chief Privacy Officers, in-house counsel and compliance professionals in the private, health public and not-for-profit sectors on managing security incidents, privacy regulatory investigations and broader data governance matters.

8. Appendix C: Members of the CANON Steering Committee

Name	Position	Organization
Adam Kardash	Co-lead	AccessPrivacy, by Osler
Khaled El Emam	CEO & Senior Scientist	Replica Analytics & CHEO Research Institute
Ruby Barber	Assistant General Counsel	Bell
Jason Brenier	Vice President of Strategy	Georgian Partners
Luk Arbuckle	Chief Methodologist	Privacy Analytics
Nicole Godin	Senior Manager, Privacy and Data Governance | Data & Analytics	Geotab
Nadine Letson	Senior Corporate Counsel	Microsoft
Faeron Trehearne	Chief Legal Officer and Corporate Secretary	Moneris
Deborah Evans	Chief Privacy Officer	Rogers
Jordan Prokopy	Partner, Privacy and Data Trust Practice Leader	PwC
Catherine Stephen	Senior Counsel	RBC
Suzanne Morin	VP, Enterprise Conduct, Data Ethics and CPO	Sun Life
Della Shea	Product & Chief Privacy Officer	Symcor Inc.

Keren Groll	Senior Counsel, Regulatory Law (Privacy)	TD Bank Group
Pam Snively	Chief Data and Trust Officer	TELUS
Noelle Paraskevopulos	Senior Legal Counsel and Chief Privacy Officer	TransUnion
Jackie Moher	Managing Counsel, Privacy & Law Clerks	CIBC

9. Acknowledgements

We wish to thank the members of the CANON Steering Committee for the effort spent on this project and the support they have provided to the project team. We also want to thank Jordan Prokopy from PwC for the significant in-kind contribution from her team to this project. CANON members also provided valuable input throughput the project formally and informally, and we wish to thank them.

10. References

[1] Innovation, Science and Economic Development Canada, “Strengthening Privacy for the Digital Age,” *Government of Canada*, 2019. https://www.ic.gc.ca/eic/site/062.nsf/eng/h_00107.html (accessed Dec. 14, 2020).

[2] Innovation, Science and Economic Development Canada, “Canada’s Digital Charter in Action: A Plan by Canadians, for Canadians,” *Government of Canada*, 2019. https://www.ic.gc.ca/eic/site/062.nsf/eng/h_00109.html (accessed Dec. 14, 2020).

[3] J. Creswell and C. Poth, *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, Fourth Edition. Thousand Oaks, CA: SAGE Publications, 2018.