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

Check against delivery

Keynote by Patricia Kosseim, Information and Privacy Commissioner of Ontario
Synthetic Data Summit
May 16, 2025

Unlocking the Value of Data While Protecting Privacy

Introduction

- Good morning. Thank you for the invitation to join you today. It's a real pleasure to be here.

- Je suis ravie de me retrouver ici à Montréal, ma ville natale qui me rappelle de si beaux souvenirs de ma jeunesse.

- This summit is a valuable opportunity to explore how synthetic data generation can be used to help solve some of the most complex societal challenges of our times.

- New and emerging innovations are changing the world in ways we never dreamed of, and at an alarmingly rapid pace.

- Data and digital technologies are now integrated into almost all aspects of our lives.

- As such, they offer a tremendous opportunity to derive the knowledge and insights we need to vastly improve our health, economy, and social well-being.

- Through technological advances, organizations can now collect vast amounts of data, harnessing its power to drive innovation and improve the delivery of goods, services, and programs.

- Data about things is usually fair game, subject to intellectual property and confidentiality issues, of course.

- But when data is about people, it gives rise to additional concerns.

- Data breaches increase the risks of washing over privacy rights and sweeping away public trust as part of the undertow.

- The challenges of accessing personal information, particularly in the health sector, have been around for a long time.

2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada M4W 1A8

2, rue Bloor Est
Bureau 1400
Toronto (Ontario)
Canada M4W 1A8

Tel/Tél : (416) 326-3333
1 (800) 387-0073
TTY/ATS : (416) 325-7539
Web : www.ipc.on.ca

- First, is the all-or-nothing nature of data identifiability. Under Canadian privacy laws, as soon as there's a reasonable prospect of identifying an individual from the data, either alone, or in combination with other data, it is considered personal information and becomes subject to the full suite of privacy protection laws with little opportunity for dialing up or down the stringency of legal requirements according to risk.

- That creates the incentive to de-identify the data, by removing direct and indirect identifiers so you can extract it from the law's application altogether, and do with it what you will. But if you don't deidentify the data properly, taking into account contextual factors, individuals may be reidentified, and you could have a significant breach on your hands. Or, conversely, if you deidentify data too well, you may reach the point where your data has lost all its utility.

- Then there's the age-old consent problem. If you have to get consent to collect and use identifiable information, it's either entirely impracticable to obtain given the mass number of people involved and the impossibility of contacting them. Or, if you can seek consent, you run into consent bias. Where, necessarily, those who agree to provide you with their consent to access and use their personal health information, already have something inherently different about them than the group that chooses not to agree, introducing bias into your data.

- Sometimes the law allows you to use personal health information without consent, subject to third-party proxy approval, like a research ethics board. But we know how difficult that can be. First to determine whether REB approval is even required. And if it is, then having to coordinate multiple boards across different jurisdictions, many of which don't have the necessary capacity or bandwidth to give you a highly sophisticated answer in a timely manner.

Enter Synthetic Data Generation

- But what if you could avoid all that by using synthetically generated data that is not about any particular individuals? Could you reap the benefits and the full utility of the data, without posing risks to privacy? That is essentially the question we are here to address today.

- Unlike all of you in this room, I am not an expert in synthetic data. But I am a strong proponent of privacy enhancing technologies (PETs) like synthetic data generation, among others.

- When I started in my role as commissioner almost five years ago, I set out with the vision of being a modern and effective regulator, with real-world impact.

- To stay true to that vision, it's incumbent on me and my staff to embrace new ideas and innovative approaches so we can support the responsible use of data in ways that can benefit our health, our economy, and our society.

- As regulators, we have to meet innovation with innovation and find pragmatic solutions to modern data challenges so we remain relevant to the organizations we oversee.

- We need fresh perspectives on methods and techniques to protect privacy across both the public and private sectors.

- A recent public opinion survey commissioned by the Office of the Privacy Commissioner of Canada found that compared to five years ago, three-quarters of Canadians are less — or much less — willing to share their personal information with organizations.

- 78 per cent have refused to provide an organization their personal information due to privacy concerns; and 41 per cent have stopped doing business with an organization that experienced a privacy breach.

- We all know that balancing innovation with privacy can be challenging.

- But privacy-enhancing technologies, like synthetic data, offer us a promising way forward, to reduce privacy risks while unlocking the potential of data.

The Power of PETs

- Last January, my office hosted our annual Data Privacy Day event. The theme was the Power of Privacy Enhancing Technologies or PETs.

- A recording of the entire event is available on the IPC's YouTube channel. If you're interested in the topic, I highly recommend listening to it.

- We were incredibly fortunate to host leading technology experts from the public and private sectors, academia, law, and privacy advocacy groups, for an in-depth and thought-provoking discussion.

- The theme of the event centered on how PETs can advance research and innovation by making it possible for organizations to use valuable data while keeping personal information safe.

- We heard about different privacy enhancing technologies including: deidentification, synthetic data generation, federated learning, and differential privacy; how each one works; what are the risks and benefits; and how they compare with one another.

- We heard real use cases of how these PETs are being applied in practice, in the public, private and health, and how organizations that truly want to succeed must stay focused on public trust as their North Star.

- And we heard how some jurisdictions are moving away from personal information as an all-or-nothing concept, towards an understanding of identifiability as a spectrum.

- Privacy law reforms either recently adopted, or being considered, in Canada and elsewhere are beginning to regulate personal information based on risk of identifiability, rather than the type of on-off switch of yesteryear that longer fits today's complex digital world.

- Dr. Khaled El Emam, the IPC's current scholar-in-residence and the key convener of today's summit, participated in our event.

- He explained how synthetic datasets made from personal information can look and feel a lot like the real thing — in terms of overall statistical properties and patterns — but they vary enough at the individual level so as not to reveal the real people they're based on. Increasingly, these artificial datasets are being generated at mass scale to train large language models and advance important health research.

- Khaled was also a guest on my Info Matters podcast for an episode about synthetic data, called *Real or Fake: The Buzz about Synthetic Data*. If you'd like to have a listen, it's episode 1 of season 2 of our Info Matters podcasts, and it's available wherever you download your podcasts.

International developments in PETs

- Worldwide, data protection regulators are supporting the adoption of privacy enhancing technologies as a reasonable way to reduce risks to privacy.

- But to really walk the talk, and help support the adoption of PETs, we as regulators need to offer concrete guidance for organizations of all sizes to help reduce uncertainty, foster trust, and inspire innovators to forge ahead.

- Fortunately, a growing number of regulators worldwide, and here at home, are actively putting forward tools and guidance to support the responsible adoption of PETs.

- In Singapore, the Infocomm Media Development Authority has guidance on synthetic data generation to help organizations understand the techniques and potential uses of this technology.

- They also have a Privacy Enhancing Technology Sandbox as a safe space to develop, test and validate innovative PETs for a limited time, under the supervision of the regulator.

- Singapore's Infocomm Authority also matches organizations with PET vendors, provides grants to implement pilot projects, and offers regulatory support to ensure compliance.

- Singapore's PET sandbox, and other regulatory sandboxes, are described in a recent research report done in collaboration with my office.

- The report is called *Exploring the Potential for a Privacy Regulatory Sandbox for Ontario*, co-authored by Dr. Teresa Scassa and Elif Nur Kumru of the University of Ottawa.

- It's available for download from our website's Research and Innovation Hub if you're interested in learning more.

- Another international initiative relevant to synthetic data generation comes from South Korea's Personal Information Protection Commission.

- The commission has issued guidance on the methods and procedures for creating and using synthetic data while remaining in compliance with privacy laws.

- And the U.K.'s Information Commissioner's Office provides guidance on PETs for data protection officers.

- It explains the different types of PETs available, including synthetic data. It recognizes that some of the vulnerabilities of real data, like bias, if not cleansed from the original data, can carry over into the synthetically generated data.

- It also acknowledges that assessing re-identification risks involved with synthetic data is an ongoing area of development.

PETs in Ontario

- Here at home, I'm particularly proud to say that Ontario was a pioneer in the field of privacy enhancing technologies.

- Back in 1995, my office introduced the concept of PETs in a joint report issued with the predecessor of the Dutch Data Protection Authority.

- The report describes how PETs can be built into the design of new information technologies to enable greater use of anonymized data for daily transactions.

- Since then, we have continued our work in this area by publishing guidance, issuing orders and decisions, and pursuing regulatory amendments and frameworks.

- For example, in 2016, we released our international, award-winning *De-identification Guidelines for Structured Data*.

- We continue to build on our expertise in this area and are in the process of updating our guidance on de-identification, taking into account the technological developments of the past decade.

- We are working with Dr. El Emam, and consulting with experts from a broad range of sectors, to seek their feedback. We expect to publish our updated de-identification guidance in the coming weeks.

- Follow us on social media for updates about the guidance and when it will be released.

- In addition to issuing guidance, our office provides advice and recommendations on government legislation, programs, privacy policies, and information management practices.

PHIPA regulation on deidentification

- In terms of law reform, Ontario's *Personal Health Information Protection Act*, otherwise known as *PHIPA*, was amended in 2020 to allow for a regulation that would eventually elaborate on the definition of "de-identify" by setting out more specific requirements.

- To this day, there's been no such regulation.

- Given the ongoing need for certainty and clarity, we took it upon ourselves to recommend specific language to the Ministry for a possible regulation on de-identifying personal health information.

- We saw this as an opportunity to provide clear rules for health information custodians to encourage more responsible data sharing.

- Specifically, we recommended that:

- The process of de-identifying personal health information must:

    - be carried out by, or under the supervision of, an individual qualified in the field with appropriate knowledge and expertise;
    - follow generally accepted best practices;
    - assess the risk of reidentification; and
    - implement steps that are reasonable in the circumstances to ensure that the risk of re-identifying an individual is very low.

- If de-identified information will be disclosed to a specific person, a written agreement must be in place prohibiting any attempt by that person to reidentify the data and requiring them to implement reasonable steps to ensure that the risk of reidentification remains very low.

- Periodic reassessments of the deidentified data must be done, taking into account certain prescribed factors, to ensure risk of reidentification remains very low.

- Certain aspects of the deidentification process must be documented, and, reasonable notice must be provided to individuals whose personal health information is being deidentified, including the general purposes for which it may used or disclosed.

- In putting forward our recommendations, we tried to ensure general consistency with a recent regulation under Quebec's, Law 25, the *Act to Modernize Legislative Provisions as Regards the Protection of Personal Information*, that sets out explicit data anonymization requirements – a first in Canada.

- We also strived to remain consistent with our policy requirements set out in our, *Manual for the Review and Approval of Prescribed Persons and Prescribed Entities*.

- These prescribed bodies include some of the largest health data custodians in Ontario that maintain registries of personal health information for the purpose of improving the provision of health care, and those that analyze and compile statistical information to better manage, evaluate, monitor and plan the health system.

- Examples include Ontario Health, ICES, Canadian Institute of Health Information (CIHI) and Cancer Care Ontario.

- Our recommendations are also consistent with the Data Integration Standards under our public sector law that enables prescribed data integration units to link data across government ministries and enable access to de-identified datasets for the purpose of:

    - managing or allocating resources;
    - planning the delivery of programs and services provided or funded by the Government of Ontario; and
    - evaluating those programs and services.

- An example is the inter-ministerial data integration unit housed in Ontario's Ministry of Health.

- Finally, we wanted to ensure our recommended regulation is also consistent with our past investigations and decisions on de-identification in the health sector, including two PHIPA-related decisions:

IPC decisions on de-identification under PHIPA

- The first one, PHIPA Decision 243, involved a group of researchers at the University of Toronto and their research database UTOPIAN, which contains de-identified patient data extracted from electronic medical records of many contributing primary care physicians.

- An anonymous complaint from a group of doctors alleged, among other things, that:

    - the de-identification process used by the researchers was inadequate, and there was a risk of disclosing potentially identifying information from the database to other researchers seeking access to the EMR data

- Our investigation found no concerns with the adequacy of the de-identification processes. However, we did recommend that:

    - the university should exercise greater transparency with custodians and open lines of communication to foster trust in their research, and

    - that they conduct a re-identification assessment based on the best practices in the IPC's guidance, *De-identification Guidelines for Structured Data*.

- We also found that the research platform continued to operate after the REB approval had lapsed, which was deeply concerning.

- Bottom line is, while research is vital to improving the quality of care and the effectiveness of our health care system, the public must be confident that their personal health information is safe and protected.

- The second case, PHIPA Decision 175, involved an investigation into the sale of de-identified data by a group of medical clinics to a third party corporation.

- We launched the investigation after becoming aware of this situation through media reports.

- There were three main takeaways from this investigation:

- First, we found that under PHIPA, the use of personal health information for the purpose of de-identifying it is permitted without the consent of the individual when specific conditions set out in the law are met.

- Second, health information custodians must be transparent about their information practices. If custodians are selling de-identified data to a third party for any purpose, including health-related research, they must clearly and explicitly reflect this in a notice to the public.

- Third, any sale agreement between a health information custodian and a third party must include adequate privacy and security controls to ensure that the de-identified data transferred to the third party remains de-identified.

Contracting with third parties

- Since PHIPA Decision 175, we have published practical contracting guidance for public institutions, including health care organizations.

- For those of you interested, I recommend our guidance, *Privacy and Access in Public Sector Contracting with Third Party Service Providers*.

- It provides practical advice to identify access and privacy considerations when contracting with third parties.

- And it includes best practices and recommendations to support proper due diligence and accountability throughout the procurement process, from planning, to tendering, vendor selection, contracting, agreement management, right up to and including agreement termination.

- As I like to repeat often, "You can outsource services or data to others, but you can't outsource accountability!"

The road ahead

- On that note, I'd like to leave with you with some closing thoughts.

- Privacy-enhancing technologies offer us promising tools to harvest the value of data while protecting privacy.

- The challenge before us is making PETs accessible, secure and easy to adopt, as a way of mitigating risks to privacy.

- PETs can no longer be seen as complex and out-of-reach technological tools that only large-scale, sophisticated organizations can afford.

- They are essential safeguards that must be accessible to all organizations — both large and small — for protecting privacy in a data-driven world.

- Both regulators and data users must rise to this challenge if we are to build the necessary trust of citizens and consumers to engage meaningfully in our digital world.

- I've spoken about doing our part as regulators.

- We must be open to novel methods and techniques to protect privacy so we can find pragmatic solutions to modern data challenges and remain relevant to the organizations we oversee.

- As regulators, we can't be naysayers all the time. We must meet the moment and meet innovation with innovation.

- It's incumbent on us to provide helpful advice and develop consistent guidance to help support responsible adoption of PETs by organizations, and give them the necessary space, certainty, and predictability they need to innovate with confidence.

- And if I may, I have a few suggestions for you as well, many of which you are already doing in this session you've convened today.

- First, as data generators and data users, it's important to explain synthetic data generation in plain and simple language people can understand. As the brilliant Einstein once put it, "If you can't explain it simply, you don't understand it well enough." Practical use cases go a long way in describing the benefits of synthetic data in concrete terms, to help demystify the process, and enable adoption of tools by others, recognizing that one size does not fit all.

- Second, be transparent and very explicit about the risks. This includes risks of identity disclosure, attribute disclosure and membership disclosure. Don't sugarcoat them. Develop metrics that quantify the level of reidentification risks and be relentless in your ongoing efforts to mitigate those risks. Again, to quote Einstein who once said of himself, "It's not that I'm so smart, it's just that I stay with problems longer."

- And third, I'd say, take up the invitation by data protection regulators to consult with them on your innovative and precedent setting approaches.

- Whether it's part of a formal regulatory sandbox, or an informal policy consultation, it's this level of engagement and collaboration that will help us better understand where each of us is coming from and help bring forth the creative and pragmatic solutions that we need.

- Through iteration and collaboration, we can find ways for organizations of all stripes to adopt PETs to help solve some of the most complex challenges we're facing in our health care system, our economy, and our society.

- Trust is the cornerstone of success. And earning that trust means being transparent and accountable.

- When people are confident that their sensitive health data is being handled responsibly, they're more willing to seek services, adopt products and contribute to research.

- In this context, compliance should be seen not as a barrier but as a catalyst for better outcomes.

- By investing in the power of PETs, we can reap the benefits of data while protecting privacy — unlocking new opportunities for all of us.

- Thank you.