INSTITUT C.D. HOWE INSTITUTE

COMMENTARY
NO. 706

The Missing Pillar of Canada's AI Strategy: Data Supply Chains

Canada has poured billions into AI talent and computing power, but weak data access threatens to hold the ecosystem back. This Commentary *makes the case for pragmatic data-sharing reforms that preserve privacy while giving Canadian firms the inputs they need to compete globally.*

Anindya Sen

About The Author

Anindya Sen
is a Fellow-in-Residence and inaugural Scholar in AI and Digital Policy at the C.D. Howe Institute, and Professor of Economics and Associate Dean of Arts at the University of Waterloo.

The C.D. Howe Institute's Commitment to Quality, Independence and Nonpartisanship

The C.D. Howe Institute's reputation for quality, integrity and nonpartisanship is its chief asset.

Its books, Commentaries and E-Briefs undergo a rigorous two-stage review by internal staff, and by outside academics and independent experts. The Institute publishes only studies that meet its standards for analytical soundness, factual accuracy and policy relevance. It subjects its review and publication process to an annual audit by external experts.

No C.D. Howe Institute publication or statement will endorse any political party, elected official or candidate for elected office. The Institute does not take corporate positions on policy matters.

As a registered Canadian charity, the C.D. Howe Institute accepts donations to further its mission from individuals, private and public organizations, and charitable foundations. It seeks support from diverse donors to ensure that no individual, organization, region or industry has or appears to have influence on its publications and activities. It accepts no donation that stipulates a predetermined result or otherwise compromises its review processes or inhibits the independence of its staff and authors. A comprehensive conflict-of-interest policy, including disclosure in any written work, applies to its staff and its authors.

Commentary No. 706
February 2026

Alexandre Laurin
Vice-President and Director of Research

ISBN 978-1-77881-073-2
ISSN 0824-8001 (print);
ISSN 1703-0765 (online)

The Missing Pillar of Canada's AI Strategy: Data Supply Chains

by Anindya Sen

- **While significant federal government investments in Canada have strengthened research capacity, talent development, and computing infrastructure, a foundational pillar of AI innovation remains underdeveloped: strong, secure, and trusted data supply chains. This *Commentary* argues that a comprehensive Canadian AI strategy must explicitly prioritize data supply chains while embedding robust protections for privacy and public trust.**
- **Current legal and institutional frameworks lack clear, standardized rules governing data sharing for AI development, and Canada does not yet have formal social-benefit tests to balance innovation gains against potential harms to privacy. These institutional gaps limit the ability of firms to access critical datasets. Coordinated policies to enable responsible access to large, high-quality datasets will give Canadian firms – particularly startups and scaleups – a strong international competitive advantage.**
- **I propose a policy framework grounded in three principles: (1) expanded, yet guarded, data access; (2) the integration of privacy-enhancing technologies; and (3) transparent governance mechanisms guided by cost-benefit analysis.**
- **Key recommendations include allowing carefully regulated private-sector access to some confidential Statistics Canada datasets on a pilot basis; accelerating the national production of synthetic datasets; embedding social-benefit tests in privacy legislation; and establishing regulated AI sandboxes aligned with regional strengths.**

Introduction

Evan Solomon, Canada's Minister of Artificial Intelligence and Digital Innovation, last fall established an AI Strategy Task Force to help define Canada's approach to AI.[1] The task force is focusing on several priorities including: research and talent; AI adoption across industry and governments; AI

1 See: https://www.canada.ca/en/innovation-science-economic-development/news/2025/09/government-of-canada-launches-ai-strategy-task-force-and-public-engagement-on-the-development-of-the-next-ai-strategy.html.

The author extends gratitude to Colin Busby, Dan Ciuriak, Susie Hendrie, Daniel Schwanen, Rosalie Wyonch, and several anonymous referees for valuable comments and suggestions. The author retains responsibility for any errors and the views expressed.

Policy Area: Industry Regulation and Competition Policy.
Related Topics: Communications and IT; Competition; Consumers' Interests and Protection; Regulatory Burden; Service Sector.
To cite this document: Sen, Anindya. 2026. *The Missing Pillar of Canada's AI Strategy: Data Supply Chains.* Commentary 706. Toronto: C.D. Howe Institute.

C.D. Howe Institute Commentary© is a periodic analysis of, and commentary on, current public policy issues. Percy Sherwood, Michael Benedict, and James Fleming edited the manuscript; Yang Zhao prepared it for publication. As with all Institute publications, the views expressed here are those of the authors and do not necessarily reflect the opinions of the Institute's members or Board of Directors. Quotation with appropriate credit is permissible.

To order this publication please contact: the C.D. Howe Institute, 110 Yonge Street, Suite 800, Toronto, ON, M5C 1T4. The full text of this publication is also available on the Institute's website at www.cdhowe.org.

commercialization; scaling of startups; developing safe AI systems that strengthen public trust in AI; facilitating skills acquisitions; and building public infrastructure. This is consistent with Solomon's earlier emphasis on scaling up the AI industry, driving adoption and trust in, and sovereignty over, the technology.[2] These strategies build on earlier federal government initiatives, such as the 2017 Pan-Canadian Artificial Intelligence Strategy (PCAIS), which was based on the three pillars of cultivating domestic research and talent, enabling commercialization, and identifying appropriate privacy standards.

On the other hand, the minister did not make any explicit mention of establishing strong data supply chains for firms, which are critical for AI innovation. In this respect, while the federal government has spent considerable effort on strengthening Canadian AI sovereignty through developing resources for domestic computing and infrastructure, less attention has been paid to improving access to harnessing large datasets. To address this gap, some experts have suggested the possibility of "data markets" where individuals could sell the information they generate through online interactions with digital platforms. Such an approach is already used by Big Tech firms to power AI algorithms for targeted advertising.[3]

However, while data markets offer an interesting possibility, concerns for privacy, the ease of data portability, and the possibility of so-called "thin markets," where there are limited buyers and sellers, have prevented the evolution of such marketplaces.[4] Instead, an attractive option would be to make certain Statistics Canada confidential datasets available (while preserving confidentiality) to private sector organizations on a pilot basis for advanced data analytics and training of machine-learning (ML) algorithms.[5] Such an approach could provide a strong boost to Canada's AI ecosystem.

While Statistics Canada does not collect nor release real-time data, its administrative datasets on individual tax records, income, employment status, and access to healthcare, for example, would be extremely difficult for private sector firms to replicate. For example, Statistics Canada's Longitudinal Administrative Databank is a 20-percent sample of all Canadian tax filers that consists of annual data for individuals on income and demographics from 1982 onward.[6] This valuable dataset links individuals through tax files and other sources. It includes rich demographic details (age, gender, marital status), income sources (wages, self-employment, social assistance, registered retirement savings plans, investments, pensions, and child tax benefits), as well as geographic location (provinces, census divisions). The databank also contains details on immigration status, including landing year, immigration category, and education level. Furthermore, these individual records are linked over time.

Another example of a rich dataset is the National Ambulatory Care Reporting System (NACRS), which contains millions of data observations on individual admissions for day surgery, outpatient and community-based clinics, and emergency departments across the entire country and over several years. NACRS contains a wealth of

2 See: https://financialpost.com/technology/new-ai-minister-says-canada-wont-over-index-on-ai-regulation.

3 Licensing the use of privately collected data through digital platforms is becoming more common. There is evidence that Reddit earns significant revenues from data licensing. See: https://www.reddit.com/r/investing/comments/1gfadkb/reddits_shares_surge_on_ai_data_licensing_deals.

4 Thin markets refer to the lack of sufficient products and/or product diversity.

5 Machine learning (ML) is a method of artificial intelligence in which algorithms learn from data to make predictions or decisions without being explicitly programmed for each task.

6 For further details, see: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4107.

information, such as the complexity of health conditions at the time of admission, time to treatment by physicians, time to admission, and demographic characteristics.[7]

These datasets should be invaluable for training ML models for a variety of private-sector purposes, such as credit risk and default prediction, fraud-detection model pre-training, labour-market forecasting, health utilization and cost prediction, and customer segmentation. Further, the attractiveness of these data are amplified by their near-universal population coverage and, hence, minimal selection bias compared to private datasets that are collected through a limited number of digital platforms. In other words, echoing Lehrer and Xie (2026), Statistics Canada datasets represent a gold-standard for training data, even if they are not collected in real time.

However, these data are based on surveys of the Canadian public, which generally trusts Statistics Canada and clearly feels that there are societal benefits from sharing its information. Hence, it is imperative that this trust be maintained through transparency about who has access. If such data are shared with the private sector, perhaps limiting it to firms that are clearly subject to enforceable Canadian privacy safeguards and, in general, implementing data governance regulations that preserve individual privacy would be sensible steps forward. This is especially important given recent studies that demonstrate how individuals can be revealed in supposedly de-identified datasets, with relatively straightforward statistical and ML methods.

Public trust can be further built through clear and transparent guidelines and new privacy legislation with social-benefit tests that require organizations to weigh the benefits and costs of data sharing, along with necessary steps to protect individual privacy. Privacy legislative frameworks in many other countries lack such social-benefit tests, and this is where Canada can innovate to ensure greater data supply to innovators while minimizing risks to individuals from privacy breaches. Sharing of data should be based on differential privacy (DP) methods, federated learning, and synthetic data.[8] Regulated sandboxes, or secure digital environments, should also be established, all of which significantly reduce the likelihood of privacy violations and should, therefore, build public trust.[9]

Beyond trust building, there also needs to be an openness toward data sharing by regulators. Studies from the Council of Canadian Academies (2015, 2023) suggest that the overly cautious approach by data custodians toward data sharing through the imposition of stringent access restrictions that require considerable effort and time to comply may impact the incentive to conduct research based on large datasets, leading to reduced innovation.[10] However, Sen et al. (2025) demonstrate that it is possible to construct guidelines for health-data sharing based on cost-benefit principles in a manner

7 Details on NACRS are available at https://www.statcan.gc.ca/en/microdata/data-centres/data/dad-nacrs-omhrs-cvsdd.

8 Synthetic data are artificially generated datasets that reflect the statistical properties of real data while reducing privacy risks.

9 Nair and Inverarity (2022) offer an intuitive explanation of federated learning methods. Conventionally, data are sent to cloud-based ML applications where the computations are executed and the trained model is made available to generate predictions or for other purposes. With federated learning, data are stored locally on a device and a ML model can be sent to that same device to be trained locally, rather than uploading the data to a less private and less secure cloud. An even greater privacy layer could be provided by using differential privacy methods where statistical "noise" is added to individual data points, making it difficult for a malicious party to infer confidential individual-level data. An example of using these methods would be to release data at certain aggregation levels rather than at the individual level.

10 Also, see https://www.oecd.org/en/blogs/2025/09/we-have-a-lot-of-valuable-health-data-why-is-it-so-hard-to-use.html. However, there are recent initiatives aimed at enhancing the sharing of health data. From 2023 onwards, patient-level data from 40 hospitals in Ontario are available to researchers through GEMINI (https://geminimedicine.ca/).

that enables data-hosting organizations to compare social benefits from innovation against possible losses in privacy and corresponding liability costs.

Currently, Canada lacks clear, standardized, and legally supported data-sharing frameworks that enable responsible AI-model training while protecting individual privacy and public trust. Without established policies, Canada risks falling behind other countries that decide to move forward in building shared data infrastructures with clear data-sharing governance principles and mechanisms.[11] Canada has an opportunity to develop infrastructure that supports data sharing within secure data access and analysis environments, enabling organizations to run algorithms without direct access to raw personal data. This must be supported by legislation on data licensing and use rights.

In this respect, I recommend that private sector organizations be given easier access to some of the vast troves of confidential data that are collected and maintained by Statistics Canada in ways that preserve confidentiality. A step forward would be a national pilot program that grants access to certain data, the results of which could be used to evaluate the benefits and costs of further expansion.

Statistics Canada datasets are among the most comprehensive in the world. Their value is amplified as they represent an incredible degree of diversity. These datasets could be invaluable for startups and other early-stage firms that critically need data to train AI/ML algorithms. This should be supplemented using other privacy-enhancing technologies (PETs) that would allow private sector organizations to access these data. In particular, the evolution of federated learning and DP methods offers intriguing opportunities for private sector innovation. In summary, enabling easier access to Statistics Canada data for private sector organizations is consistent with establishing a strong data supply chain.

Intuitively, secure data supply chains for AL/ML algorithms should be: (1) statistically representative and follow well-defined collection standards; (2) stored securely; and (3) available to non-government organizations under strict privacy protocols and be based, when appropriate, on the use of PETs. Statistics Canada is well-positioned to enforce all these requirements. The federal government can ensure additional data sources by prioritizing the production of synthetic, artificially generated data. This could be facilitated through appropriate regulations. The federal government should also consider further funding Mitacs, a national non-profit that brings academics and businesses together, for proposals using Statistics Canada confidential data.[12]

Canada should also learn from the UK, which is implementing regulated sandboxes that enable firms to test the impacts of AI technologies on actual market participants, but under supervision. Separate sandboxes should be established for different sectors in the economy and built on regional strengths. Furthermore, they should be implemented in a manner that encourages research collaboration between universities and local firms. This would lead to greater diversification in AI innovation and ensure that the benefits of such innovation are spread across the country.

11 An example is the Genesis Mission, which is the US government's strategy to build an integrated AI platform to harness federal scientific datasets to train scientific foundation models and create AI agents for a variety of purposes (https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/). China has also moved forward transparently in specific AI governance strategies. See: https://www.mayerbrown.com/en/insights/publications/2025/10/artificial-intelligence-a-brave-new-world-china-formulates-new-ai-global--governance-action-plan-and-issues-draft-ethics-rules-and-ai-labelling-rules

12 For more information, see https://www.mitacs.ca/about/mitacs-strategic-plan/.

Past Federal Initiatives

While arguments can be made that the amount of funding provided by the federal government through its PCAIS was not comparable to that of some other countries, its vision was clear, and initiatives such as the establishment of national AI institutes have borne substantial fruit.[13] The Canadian Institute for Advanced Research (CIFAR) was asked to lead the PCAIS, which achieved several successes, including the establishment of national AI institutes in Edmonton (Amii), Montreal (Mila), and Toronto (Vector). Using federal funding, these institutes have led and continue to lead initiatives to stimulate foundational AI research, as well as connecting the research to potential commercial applications and assisting in the adoption of these new technologies.[14]

CIFAR was also responsible for developing initiatives to attract and retain global research talent (for example, through endowed research chairs) and promoting training and knowledge mobilization programs. Another noteworthy program through PCAIS was $40 million in funding provided in Budget 2021 (from 2022/23 to 2026/27) to AI researchers across Canada.

Meanwhile, the 2024 budget acknowledged the need for measures and investments that were necessary to make Canada a global leader in innovation, which could be accomplished by leveraging the benefits in the country's existing AI ecosystem – from development to commercialization to safety.[15] The announced initiatives revolved around the pillars of: (1) enhanced access to public computing resources; (2) specific measures for small- and medium-sized businesses; (3) a strong environment to encourage AI startups; and (4) facilitating skills acquisition. For example, under that budget's Canadian Sovereign AI Compute Strategy – intended to stimulate the development of AI public infrastructure – $2 billion was specifically earmarked to build and provide access to computing capabilities and technological infrastructure for AI researchers, startup businesses, and organizations interested in scaling up operations. Some of these funds ($240 million) were given specifically to Toronto-based AI consultancy Cohere to build an AI data centre that would enable it and other Canadian firms to access computing capacity.[16]

Parallel to this initiative, the government launched a $300 million AI Compute Access Fund to provide matching financial resources for small- and medium-sized businesses interested in purchasing cloud and non-cloud-based AI services. In addition, $200 million was earmarked to encourage the launch of AI-based startups in critical sectors such as agriculture, clean technology, healthcare, and manufacturing, to be disbursed through the country's regional development agencies.

13 For example, the UK has committed two billion pounds to AI innovation, development, and adoption through its AI Opportunities Action Plan (https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan).

14 In Budget 2021, each institute was eligible to receive up to $20 million from 2021/22 to 2025/26. Budget 2021 also put aside $125 million to set up global innovation clusters in Digital Technology, Protein Industries Canada, Next Generation Manufacturing Canada, Scale AI, and Canada's Ocean Supercluster. The objective of the clusters is to promote made-in-Canada AI technologies, hence stimulating Canadian innovation. For further details, see: https://www.canada.ca/en/innovation-science-economic-development/news/2022/06/government-of-canada-launches-second-phase-of-the-pan-canadian-artificial-intelligence-strategy.html.

15 Further details are available at https://www.pm.gc.ca/en/news/news-releases/2024/04/07/securing-canadas-ai.

16 See https://www.canada.ca/en/department-finance/news/2024/12/deputy-prime-minister-announces-240-million-for-cohere-to-scale-up-ai-compute-capacity.html. The federal government has issued a call for tenders for the development of more infrastructure (https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy).

Facilitating new skills acquisition was not ignored, as $50 million was devoted to the Sectoral Workforce Solutions Program designed to provide new skills training for workers in potentially disrupted sectors and communities. Finally, allocating $50 million toward establishing the Canadian AI Safety Institute is a positive step forward in promoting AI research that is aligned with mitigating possible harm while considering societal benefits.

Broadly speaking, it is fair to say that the federal government has made some sound investments in strengthening Canada's AI research ecosystem, beginning the process of domestic capacity in computing infrastructure, facilitating access to computing resources for small- and medium-sized businesses, and initiating skills development programs. The 2025/26 federal budget builds on previous spending by earmarking $925.6 million over the next five years to support the development of "sovereign" public AI infrastructure through enhanced computing availability and capacity for private and public research.[17] In terms of a new initiative, the federal budget also proposes $1.7 billion to attract international AI researchers.[18] On the other hand, the budget did not propose any new data-sharing or liberation strategy.

Privacy and Data Sharing

Implications for Privacy

AI systems are fueled by data, so it is critical that Canadian firms can access large datasets for building ML algorithms while protecting individual privacy. In other words, for an AI-based firm to flourish, it must have access to secure data supply chains. An example will illustrate how better data access can result in improved ML algorithms and, therefore, significant innovation with strong social benefits. As discussed, the NACRS contains millions of observations with comprehensive individual patient characteristics (e.g., age, sex), reason for visit, diagnosis codes, clinical severity indicators (e.g., triage level), time stamps (e.g., arrival time, time to physician assessment, discharge), and patient outcomes (e.g., discharged, admitted, transferred, death).

From a patient-benefit perspective, a ML model can statistically link inputs (patient demographics, severity of condition, time-to-physician assessment, etc.) to outcomes (mortality and risk of clinical deterioration while waiting), by learning from existing data patterns. The ML algorithm can also identify the relative importance of each of these inputs with respect to different outcomes. The results could then be used to support clinical decisionmaking by estimating risk levels for patients in ERs, improving system-level planning by identifying bottlenecks where wait times have the largest impact on patient outcomes, informing policy and resource allocation by quantifying how delays translate into measurable health risks, and providing guidance on appropriate investments in emergency-care capacity.

To construct robust supply chains, it is important to recognize that data have particular characteristics, such that they are nonrival. In other words, the use of data by one firm does not preclude its use by others, making it, in a sense, a non-exhaustible resource. However, this does not imply that data can be freely shared without consequences. Many studies have shown how access to seemingly disparate datasets could lead to a significant increase in the likelihood of reidentification of individuals,

17 See: https://www.cbc.ca/news/politics/federal-budget-quantum-ai-computing-9.6966549.
18 See: https://universityaffairs.ca/news/post-secondary-leaders-express-mixed-reactions-to-budget-2025/.

compromising their privacy.[19] As noted by Culnane et al. (2017), reidentification occurs when it is possible to identify "digital fingerprints" across datasets. Digital fingerprints are a combination of features that uniquely identify a person, which can then be used to link an individual across datasets. Hence, even if an individual's name has not been revealed in a dataset, a digital fingerprint could reveal the person's identity if another dataset has the person's name.

Culnane et al. (2017) studied the consequences of the Australian government's policy of open government data, which led to its federal health department publishing online the de-identified longitudinal medical billing records of 10 percent of Australians (roughly 2.9 million individuals). Those records demonstrate that it was possible to recover some individuals without decryption by combining the published data with other external information sources. For example, the study was able to reidentify professional athletes by matching the dates and types of treatments with available online information on injuries and surgeries. There are many other studies which establish the possibility of reidentification without decryption. Hence, any public sharing of data must take this possibility into account.

Perhaps more relevant are findings from Abowd et al. (2025), who demonstrate that it is possible to reidentify individuals in the 2010 US Census by combining publicly available microdata with individual-level information and aggregated data containing geographic identifiers. The study also establishes that ML-based differential privacy methods adopted for the 2020 US Census better preserved privacy than other methods such as record-swapping algorithms. Such methods also result in less degradation of data quality than other feasible alternatives.

The idea of differential privacy was initially put forward by Dwork et al. (2006). It is a framework based on evaluating the likelihood of compromises to individual privacy by understanding the relative importance of individual observations. In terms of implementation, DP methods inject a certain amount of noise in a data request that prevents the identification of individuals even if the information is subsequently recombined with other third-party datasets. DP-based disclosure control may recommend a combination of strategies, including aggregation, suppression of small cells, and the addition of calibrated statistical noise. These methods are currently being studied by Statistics Canada for potential use in the Canadian census.[20]

For example, suppose that Statistics Canada finds that releasing data on the number of residents in neighbourhoods by immigrant and religious status leads to small cell sizes, increasing the possibility of reidentification. Rather than releasing neighbourhood-level counts of residents by religion – where small cell sizes could increase reidentification risk through linkage attacks – data may be aggregated to higher geographic levels or subjected to DP-based noise injection to preserve privacy. On the other hand, while these approaches reduce identification risk, they may also reduce analytical precision and limit the utility of the data for fine-grained policy analysis. Despite this shortcoming, any extensive release of Statistics

19 The well-known case of data de-identification based on digital fingerprints occurred when Netflix published a dataset of more than 100 million user movie ratings from individual customers in October 2006, part of an open competition to develop algorithms to best predict user ratings. While the dataset did not have the identity of Netflix subscribers, Narayan and Shmatikov (2008) demonstrated that specific customers could be reidentified by matching the user ratings with corresponding ratings from a public website of movie ratings (IMDb) that had attached names. Individual privacy was thus compromised, as users may have rated sensitive movies on Netflix, assuming privacy, and not on IMDb. One user reidentified in this way had rated sensitive movies privately on Netflix, but not on IMDb.

20 See https://www150.statcan.gc.ca/n1/pub/12-206-x/2025001/04-eng.htm for further details.

Canada confidential data should evaluate the use of DP methods including any other statistical methods that are based on preserving individual privacy. Besides mitigating harm from a loss in privacy, such regulations will assist in encouraging public trust.

The above discussion suggests that privacy legislation must evolve to account for advances in ML/AI methods that can reveal individuals in datasets, even after de-identification. However, data sharing must still be permitted to allow AI innovation. The challenge is to develop a social-benefit framework that allows data sharing while mitigating risks from privacy loss. In this respect, different jurisdictions adopt varying approaches based on anonymization and de-identification. Under both the EU General Data Protection Regulation (GDPR) and the US *Health Insurance Portability and Accountability Act* (HIPAA), data may be shared or processed if individuals are anonymized and cannot therefore be reasonably identified. The EU defines personal data as information relating to an identified or identifiable person and excludes anonymized data where identification is "no longer possible," after factoring "all the means reasonably likely to be used… by the controller or by another person" (GDPR Recital 26; Art. 4[1]).

HIPAA similarly provides that health information is no longer protected once de-identified so that there is "no reasonable basis to believe" it can identify an individual (45 CFR §164.514[a]). In contrast, the UK Information Commissioner's Office explicitly recognizes that anonymized data can become identifiable "when combined with other information" and, therefore, requires a risk-based approach to anonymization.

Recent attempts to review privacy legislation applicable to the private sector in Canada include Bill C-11 (*Digital Charter Implementation Act*, 2020) and Bill C-27 (*Digital Charter Implementation Act*, 2022). However, these bills were not passed, and the *Personal Information Protection and Electronic Documents Act* (PIPEDA) still regulates matters relating to privacy.[21] PIPEDA and related provincial laws (e.g., Ontario's *Personal Health Information Protection Act* and Alberta's *Health Information Act*) take a similar approach: information that is de-identified or anonymized falls outside the definition of "personal information" and thus outside the Acts' scope.[22] In this respect, it is important to note how the terms "anonymization" and "de-identification" have been in many cases used interchangeably.

However, these terms have different implications. Schwanen (2023) defines anonymization as ensuring that no individual can be identified from the information, whether directly or indirectly, by any means. In other words, anonymization is irreversible. On the other hand, de-identified data implies that while personal information does not directly identify the individual, their identity could still be revealed, for example, using external datasets. This is consistent with interpretations by federal and provincial privacy commissioners in Canada. For instance, the Office of the Privacy Commissioner of Canada (OPC) notes that "de-identified information may still pose a risk

21 Both these bills contained different Acts. For example, in 2022, Bill C-27 was tabled as the *Digital Charter Implementation Act*, and it proposed to enact the *Consumer Privacy Protection Act* (CPPA), the *Personal Information and Data Protection Tribunal Act* (PIDPTA), and the *Artificial Intelligence and Data Act* (AIDA). The bill died on the order paper in January 2025 when Parliament was prorogued.

22 However, as noted by Scassa (2022a), PIPEDA does not explicitly define "de-identification," nor does it contain specific rules around the use or disclosure of de-identified information.

of reidentification when combined with other datasets," and that organizations must apply reasonable measures to mitigate this risk.[23]

Another key point is that none of the above jurisdictions have explicit social-benefit or cost-benefit frameworks for data sharing, although the UK acknowledges the public and social benefits of responsible data sharing as part of good governance.[24] While Canadian law does not include a formal social-benefit test for data sharing, the OPC and recent legislative proposals, such as Bill C-27, acknowledge that the use of de-identified data for research, innovation, or socially beneficial purposes may be permissible within a regulated framework. Having a clear and transparent social-benefit data-sharing framework is critical for building public trust, and moving toward greater data sharing with private sector organizations is an area in which Canada can develop innovative legislation.

In this respect, Sen et al. (2023) demonstrate that such frameworks or tests are possible for health data and consistent with traditional cost-benefit tests applied by the Treasury Board of Canada. Canada's next round of privacy legislation must include explicit tests to enable wider data sharing of public datasets. Bernier (2021) notes that the former Bill C-11 allowed for the "disclosure of de-identified data without knowledge or consent only to specified public institutions or an organization mandated by such an institution, and for a 'socially beneficial purpose.'" She notes that this raises the question of why use by private sector organizations, which are the engines of innovation and economic growth, would never be "socially beneficial."

Future privacy legislation must enable wider sharing of de-identified data for AI training. There are mechanisms such as regulatory sandboxes and federated learning methods that can appropriately minimize risks of privacy loss and harm to individuals. As in other jurisdictions, policymakers should provide explicit guidance on best practices for data anonymization and de-identification.[25] For example, the US Department of Health and Human Services has guidelines for de-identification of data in accordance with the HIPAA Privacy Rule.[26] Further, Ontario has also recently issued guidance on de-identification.[27]

Hesitation in Data Sharing

The issue of appropriately preserving individual privacy is tied to a lack of trust on the part of data custodians, which leads to reduced data sharing and arguably less fuel for AI model training. For example, as noted by Sen et al. (2025), despite risk mitigations or safeguards in place and social benefits, many health organizations are still hesitant to share data for research, given unclear regulations and liabilities. A common argument is that limited access to confidential data is more ethical as it reduces the likelihood of data breaches and preserves individual privacy.

Data custodians also have an incentive to be risk-averse if they are potentially liable for breaches of individual privacy. On the other hand, CCA (2015) notes that a failure to share data might be a significant social opportunity cost if it prevents research from being conducted to benefit society, particularly disadvantaged communities. In such a scenario, the failure to share data could be interpreted as being unethical (Stanley and Meslin 2007). Additionally, CCA (2023) highlights the prevalent overly cautious approach toward health-data sharing

23 Refer to Bernier (2021) for more details.

24 See Bernier (2021) for further discussion.

25 Scassa (2022a) notes that this guidance on best practices was significantly missing in Bill C-27.

26 The guidelines can be accessed at https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html.

27 See: https://www.ipc.on.ca/en/resources/de-identification-guidelines-structured-data.

in Canada and argues that the societal benefit of data is significantly greater when it is shared, rather than being held in siloed information systems.

The Organisation for Economic Co-operation and Development (OECD) Recommendation on Health Data Governance (2016) encourages countries to make health data available for public-interest uses such as research and innovation, while safeguarding individual privacy.[28] The document notes that Australia, Finland, France, Germany, and the UK have introduced legal frameworks that explicitly allow the use of health data for "public interest" purposes without consent but with enhanced privacy safeguards in place. This perspective reflects an important shift in public health-data governance. Girott et al. (2025) note that fragmented data governance approaches, unclear, complicated, and slow approval processes for access to health data, along with a lack of public trust, have all been significant impediments to the sharing of health data for innovation purposes. The establishment of the European Health Data Space (discussed below) is a step forward toward encouraging data sharing. However, while different countries have been more explicit about incorporating public-interest purposes for data sharing, there is still no specific social-benefit test.

Establishing Cost-Benefit Guidelines for Data Sharing

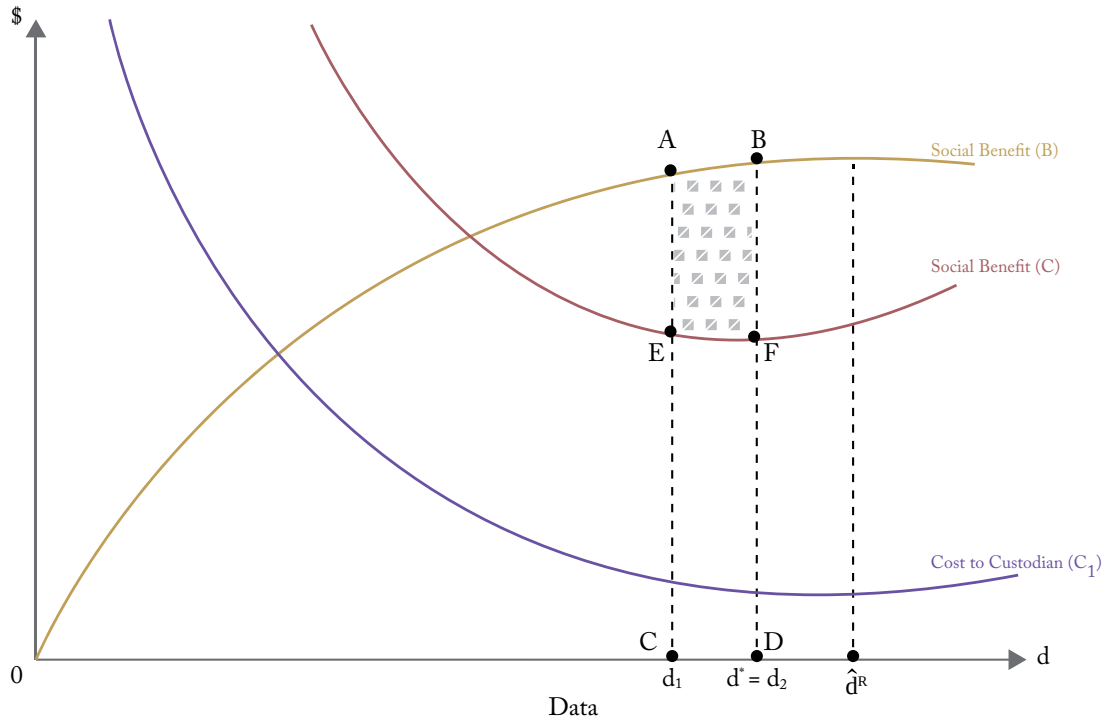
Given the absence of explicit social-benefit tests for data sharing, it would be useful to outline a possible approach. One is the model proposed by Sen et al. (2025), which could be employed by custodians of health data, such as hospitals that maintain individual-level patient data. The data custodian, typically a research ethics board, has authority with respect to granting researchers access to data (*d*). The data custodian does not gain any direct benefits from allowing access to data, but experiences costs (*C*) in establishing infrastructure that stores data and enables access by researchers. Public or private sector organizations are interested in accessing the data (*d*) to run ML/AI algorithms that will eventually result in some benefit (*B*) to society through the creation of new knowledge and/or innovative products/services.

Data custodians need a social-benefit test as they must evaluate if the societal benefits from knowledge creation are maximized while minimizing the probability of individual privacy being compromised. This is because the trade-off to knowledge creation from allowing access to data is the probability that individual-specific information contained in the database may be revealed, despite the use of privacy-preserving technology by the data custodian. The cost of employing these technologies is also captured by *C*. But these are not the only costs that might be incurred from societal data sharing. Sen et al. (2025) note that if an individual's privacy is compromised, then we must assume that they will experience some harm (*H*), which can be monetary or non-monetary.

The variable *H* captures both the probability of being harmed, as well as the actual monetary and non-monetary amount of harm, and reflects the costs to individuals, which might occur from reidentification through, for example, combining datasets permitted for analysis with other external information. Sen et al. (2025) assume that sharing more data leads to a higher probability of data being accessed by unauthorized third parties and, therefore, an increase in harm (*H*) to individuals, which grows at an increasing rate. With respect to the provision of data, if the incremental cost of providing it to another user is very low, then the data custodian experiences economies of scale in supplying data with some initial fixed costs for maintaining data infrastructure and the employment of data protection technologies. The societal costs of providing data to public and private

28 See: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0433.

Figure 1: Basic Cost-Benefit Model



Source: Authors' conceptualization.

bodies are then the sum of the data custodian's operations costs and the possible harm costs to individuals.

What is the implication of this model? While the societal cost curve will initially fall because of the economies of scale associated with data provision, it will start rising as more data sharing implies increasing harm. The gap between the private cost curve of the data custodian and the societal cost curve is the amount of harm to individuals from privacy breaches.

Finally, there are social benefits from knowledge creation and innovation that are an increasing function of data access, that is, of each data unit. In terms of notation, this is $B = B(d)$. However, there is a plateau, beyond which no further benefits are created, even though access to more data continues. This could conceivably occur when, at some point, more data does not necessarily translate to more information. Furthermore, additional data could be of lower quality after some threshold. Hence, while each unit of data initially yields a higher incremental benefit, at some point there is no positive marginal benefit. Intuitively, it is possible that additional data to train ML algorithms do not yield any incremental benefit once a dataset reaches a certain size. The optimal amount of data release from a societal perspective is then defined as the point at which the slopes of the marginal benefit and the marginal cost (to the data custodian) are equal. Figure 1 shows the basic equilibrium, with the societal benefit and cost curves, along with the cost curve for the data custodian.

This model provides a framework for estimating

net social benefits from data sharing while accounting for data provision costs and possible harm from privacy loss. The important lesson is that the data custodian must compare the marginal benefits of data release against the corresponding marginal social costs. These benefits and costs can be calculated using the corresponding areas beneath the curves with respect to the amount of data being released. For example, suppose the data custodian receives a request for an additional amount of data equal to $d_2 - d_1$. In this case, the marginal benefit to society is given by area ABCD. The corresponding marginal social cost is CDEF. As society will receive a net marginal benefit of ABEF, it makes sense for the data custodian to release the data.

The privately optimal amount of data release to the data custodian, if they are considering only their own costs of maintenance and release, is given by $d = \hat{d}^R$. On the other hand, taking into consideration harm to individuals, the optimal amount of data release is given by $d = d_2$, where the slope of the societal benefit and cost curves are equal. In other words, the socially optimal amount of data sharing will be lower because it incorporates the harm to individuals from privacy loss associated with more data sharing. The assumptions built into the model imply that there will always be some loss in privacy from data sharing. The key point is to mitigate this risk and then balance it against corresponding social benefits. Technologies such as federated learning and the use of synthetic data can significantly minimize the possibility of privacy loss and associated harm to individuals.[29]

The model is relevant as it fits the Unity Health Toronto network's GEMINI project, which is making hospital-level data available for research and, hence, socially beneficial purposes while incurring the resources to meet regulatory standards. These data and more Ontario patient-level datasets are available through the Institute for Clinical Evaluative Sciences (ICES), which is funded by the Ontario Ministry of Health.[30] However, de-identified patient-level data are typically not provided directly to private sector firms. Since 2016, ICES has allowed private sector researchers to use its data and analytics services division to obtain analyses for approved projects, where analyses are performed by ICES staff and scientists (Schull et al. 2017).

The above discussion implies that, for data sharing to succeed, having formal data sharing mechanisms with clear data governance principles based on privacy and trust is important. This is critical in the context of recent incidents, which have raised public concerns about apparent unauthorized access to individual-level data, despite the existence of public-good elements. As noted by Scassa (2022b), one example of such concern is the 2018 Statistics Canada request to credit agency TransUnion for individual-level data financial transactions and credit histories, without individual prior consent.[31] Another example is the significant media coverage of the Public Health Agency of Canada's 2022 use of de-identified data aggregated across individual TELUS cellphone subscribers to study patterns in the spread of COVID-19. Sensationalistic headlines gave the impression that the federal government was violating the individual privacy of Canadians by attempting to monitor the movements of individuals and using cellphone-

29 There are clear limitations to the above model. For example, the optimal amount of data and the potential for breaches should also be related to the context of AI development, such as the algorithm's type and purpose.

30 See https://www.ices.on.ca/ for further details.

31 See: https://www.forbes.com/sites/cognitiveworld/2018/11/04/canadians-up-in-arms-privacy-without-consent-and-the-dangerous-precedent/ for further details.

based mobility data, which they did not have permission to access.[32]

However, using cellphone-generated location data to understand patterns in disease movements undoubtedly has significant public health benefits in terms of efficient policy responses, such as guiding resources to areas with high levels of COVID-19 incidence. The data were reidentification and aggregated, minimizing the risk of revealing individual-specific information. It is also important to note that Google aggregated individual mobility data collected from individual cellphones through its Google Maps app and made the data publicly available. Assurances on privacy preservation were based on data being aggregated for sufficiently large geographic areas that would make de-identification of individuals extremely difficult, if not impossible.[33]

While much of the media coverage regarding the use of TELUS data was inaccurate, a positive spillover was an inquiry into the matter and a subsequent report by the House of Commons Standing Committee on Access to Information, Privacy, and Ethics. It made several important recommendations regarding modernizing Canada's privacy laws, improving the processes for the federal government's use of privately collected data, and improving population data literacy. The key point is that a transparent, well-established data-sharing cost-benefit framework would likely have mitigated many of these concerns.

There is also a critical need to protect intellectual property. While OpenAI has been an innovative disruptor, it is facing lawsuits from Canadian and US media companies for allegedly infringing their intellectual property by using their content without permission. Cohere is also facing lawsuits for illegal data scraping. Unresolved copyright questions around AI training data demonstrate the uncertainty in Canadian copyright and privacy legislation, which further undermines public trust. In this respect, as noted by Lehrer and Xie (2025), most data used by firms to train AI systems are proprietary, and hence there are limited incentives for these firms to share their data, resulting in significant barriers to entry for smaller firms, startups, and university researchers. In summary, there is a need for data-sharing mechanisms that can lead to greater, affordable access to large datasets.

Data Markets

The implicit assumption of the previous section is that data are usually accessed through some data custodian, possibly in a quasi-governmental institutional setting. However, there has been increasing discussion – and, in some cases, practical movement – toward the development of data markets in which individuals and organizations can purchase and sell information without an intermediary data custodian or formal privacy protection. Sen (2022) examines the efficiency of data portability regimes, under which individuals have the right to access and sell their personal data. In the European Union, the General Data Protection Regulation mandates that individuals have the right to receive their personal data from data controllers in a structured, commonly used, and machine-readable format. Currently, there is no federal legislation in Canada that mandates data portability,[34] although Quebec has such legislation.[35] Despite these legislative shortcomings, Canadians can already download engagement-created information from many social media platforms and Internet browsers.

32 See: https://nationalpost.com/news/canada/canadas-public-health-agency-admits-it-tracked-33-million-mobile-devices-during-lockdown and https://globalnews.ca/news/8503895/watchdog-probing-officials-cell-location-data/.

33 The data are available through Google's mobility reports at https://www.google.com/covid19/mobility/.

34 This was to have been a feature of Bill C-27.

35 See Kashdaran (2024) for details.

Sen (2022) assumes that individuals share ownership of the data they generate through online engagement with multi-sided platforms and can sell or transfer it under data portability rights. Multi-sided platforms (MSPs) are firms that offer free services and, in return, collect information generated by individuals through their online engagement. These data can be used to develop algorithms for targeted advertising. The best examples of MSPs are, of course, Google and Meta.

Sen does not propose that MSPs directly compensate users for employing their information in targeted advertising. Instead, he suggests that the data generated by users through MSPs should be easily portable, allowing platform members to send their information to other platforms and/or firms for compensation through secure application programming interfaces.[36] In this sense, individuals should have property rights – shared with the organization with which they engage – over the content that is generated through their engagement.

However, facilitating data portability does not guarantee that there will be a large number of purchasers for individual-level data and, in fact, data purchasers might be interested only in information on high-income individuals for targeted advertising or marketing purposes (Savona 2020). Researchers are divided on whether data markets can be successful. Lanier and Weyl (2018) and Arrieta-Ibarra et al. (2018) argue that data markets could redistribute economic surplus and spur innovation. Others, such as Acemoglu et al. (2021), caution that individual data sales can impose harms on others, since individuals who choose not to sell their personal information may nevertheless experience a loss of privacy if other individuals with similar characteristics do sell their data through data markets.

On the other hand, Jones and Tonetti (2020) suggest that there are societal welfare gains when individuals are able to choose how much data to sell, balancing privacy costs and societal benefits. This could occur if firms that have access to large datasets have an incentive to hoard the data. However, enabling individuals to directly sell their data through markets could eventually lead to enormous amounts of information freely circulating, which leads to an increased likelihood of reduced privacy. An extreme, although not implausible, scenario is individuals becoming more susceptible to cyberattacks or targeted manipulation by foreign governments.

Data Sharing in Other Countries

Data markets have begun to emerge in other countries. As detailed by Lehrer and Xie (2025), the US stands out in terms of having private data marketplaces, such as the Amazon Web Services (AWS) data exchange. For example, AWS enable subscribers to access third-party data (e.g., financial, satellite, health) that can be directly integrated into their workflows. There are also a growing number of firms offering synthetic data for sale that are relevant to finance, marketing, and healthcare professionals.[37] These digital platforms rely on self-regulation to develop user licenses and terms of service for privacy protection. However, there are no large-scale US markets that enable individuals to sell the data they generate through online engagement.

36 Open banking enables consumers to port and share their financial data across different financial institutions through secure application programming interfaces. This is the foundation of how data markets may work in practice. Open banking is evolving in Australia, the UK, the EU and Japan but is not progressing in Canada.

37 See: https://solomonpartners.com/2025/09/08/synthetic-data-is-transforming-market-research/ and https://www.jpmorganchase.com/about/technology/research/ai/synthetic-data.

Data markets in China operate more as data exchanges with government regulation. Lehrer and Xie (2025) note that China's approach is state-directed and strategically coordinated, as data exchanges are treated as national digital infrastructure and linked to priorities in AI development, smart cities, and fintech regulation. Data exchanges in Shanghai, Shenzhen, Beijing, and Guangzhou offer real-world and standardized training datasets, where buyers purchase usage rights rather than ownership.[38] Furthermore, they often use secure computing environments (trusted sandboxes) that allow analysis without access to raw personal data, thereby protecting individual privacy. However, these privacy-preserving features operate within a governance framework that affords the state broad access and oversight powers, and they should therefore be understood as protecting against unauthorized private use rather than limiting state surveillance.

Some initiatives in Europe seem aimed at facilitating data sharing through data intermediaries as opposed to sale through data markets. A good example is the European Health Data Space, which aims to create federated data infrastructures where health data are shared through common standards and are not for commercial sale. European Union regulations in May 2025 mandated timelines for member states to establish national digital health authorities to oversee data-sharing implementation, which includes certification of healthcare provider and vendor systems for interoperability and security compliance.[39] The objective is to provide EU citizens with better control over their personal health data and to ensure that various stakeholders such as researchers, companies, and policymakers can apply for access to health data for secondary purposes, which would be beneficial for training AI algorithms.

Opening Statistics Canada Data to Non-Government Organizations

Canada can become a global leader in allowing private sector organizations to access confidential data collected by different levels of government. Currently, such confidential data – for example, individual tax records, census information, individual admissions to emergency rooms in hospitals across the country and so forth – are primarily available to either university researchers or researchers affiliated with government organizations. Some non-governmental and private sector organizations can also access the data.[40]

The Canadian Research Data Centre Network's Access and Fee-for-Service policy explicitly states that "all researchers working for the private sector including industry associations" are subject to fee-for-service charges to access Statistics Canada microdata.[41] However, while it seems that Statistics Canada microdata are available to private sector researchers, I could not find strong evidence of their use by non-governmental organizations. Nor could I find any documentation of why such organizations typically do not use these data. Easier access to Statistics Canada microdata at reasonable rates would strengthen the data supply chain, which could be critical for startups that need to train ML/AL algorithms.

Furthermore, Canadian businesses could gain a comparative advantage since the UK and other European countries do not currently have such data-sharing arrangements available to researchers outside academic institutions. Of course, strong

38 There is variation in the type of data offered across data exchanges. For example, Shanghai's exchange focuses on financial and industrial data, whereas Shenzhen's offers data used in smart-city applications and industry ecosystems.

39 See: https://www.ey.com/en_gr/technical/tax/tax-alerts/regulation-2025-327-establishing-ehds for further details.

40 See: https://www.statcan.gc.ca/en/microdata.

41 See: https://crdcn.ca/app/uploads/2021/04/crdcn_affs_policy.pdf.

protocols must be implemented to ensure that greater access to confidential datasets does not compromise their quality or the privacy of individuals through adversarial ML cyberattacks that result in data breaches.

To use confidential data from Statistics Canada, researchers must first explain the reasons for requesting access to specific datasets and why their research could contribute to existing literature.[42] If the proposal is deemed acceptable, research applicants must still undergo criminal checks to further ensure the security of data access.

Once researchers successfully pass these hurdles, they are able to access the data only through specific research data centres in universities across the country. The datasets that the researcher wishes to study are first sent to one of these centres. The researcher is then able to study the data but is not able to transfer the data. Furthermore, they can only use the statistical packages and computers housed in the centre. Finally, any research output that they wish to use must first be vetted by a centre representative to ensure that the confidentiality of individual survey participants is not compromised.

This framework could be leveraged to allow data access by innovative firms. As in the academic process, they would have to justify the need to access the data and explain why suitable alternatives were not located. The evaluation committee would then decide whether granting data access may result in societal benefits. The metrics in this case would be associated with the firm's commercial success and the likely impacts on the Canadian innovation landscape. In spirit, this process is consistent with the data-sharing, cost-benefit approach that was suggested earlier.

However, it should also be emphasized that the private sector would likely view this process as burdensome. While access to data would be theoretically possible, the above protocols might still not be attractive enough for firms to participate, given the potentially large amount of time needed to evaluate and approve applications and the perceived lack of transparency in being given access to the data.

In this respect, it is relevant to note the differences in data-access cultures for academic researchers between Canada and the Nordic countries that have existed for many years. Specifically, being allowed only to study confidential microdata through research data centres makes it more difficult for Canadian researchers to engage in their work, given that the centres are typically open only during regular office hours when academics also have teaching and administrative responsibilities. Researchers may then only be able to conduct their research by employing research assistants, which has clear benefits to students but also becomes costly. In contrast, academics in most Nordic countries are allowed to access confidential administrative data across countries through their secure desktops and virtual private networks (VPNs).[43] The good news is that steps have been implemented to establish virtual Research Data Centres (vRDCs), which will allow Canadian academics to finally access specific confidential datasets from work or home. While this type of data access is welcome, it should not have taken this much time for Canada to emulate the type of remote access that is available to researchers in Nordic countries.

If private sector organizations are granted easier access to confidential microdata, there are privacy-enhancing technologies that can protect individual privacy. The use of such technologies would further allay public concerns about allowing private sector organizations to access confidential data. One obvious approach would be to ensure that only encrypted data are shared, with encryption keys being sent to a limited number of designated

42 Details are available at https://www.statcan.gc.ca/en/microdata/data-centres/access.

43 For example, see Statistics Denmark (2014) and Thalow and Nielsen (2015).

researchers. Another strategy would be to consider DP methods before data release. However, the loss of accuracy that accompanies DP may reduce the power of ML algorithms that private sector organizations might want to develop.

Another potential initiative would be to offer federated learning services[44] in which a central organization could run ML models using data from different distributed sources. This leads to the possibility of a private sector organization requesting Statistics Canada to run different ML models using their data along with confidential data held by Statistics Canada. The private sector firm would not be able to access the confidential data collected by Statistics Canada and would instead receive relevant "weights" that can be used in its algorithms.

The federal government should consider specific mechanisms to encourage businesses to use Statistics Canada data. A possible path would be a national program through Mitacs that funds the hiring of students. However, in this case, the proposals should be for larger sums of money ($75,000 and more) that enable the hiring of graduate/advanced undergraduate students or recent graduates. Furthermore, as with established Mitacs policy, these grants should fund academic and business partnerships. This type of program would strengthen the strong research relationships that exist between universities and businesses, as well as facilitate a richer AI data ecosystem.

The UK approach is another example of a national-level strategy intended to stimulate AI innovation by private sector firms. For example, the UK government recently announced the launch of the AI Growth Lab that will allow firms to test the impacts of their AI technologies in regulated sandboxes under government supervision to evaluate possible real-world impacts.[45] The regulated sandbox that the UK is considering would enable firms to test technologies with actual market participants. One example is the possibility of testing the capability of AI-based computer vision technologies to help radiologists.

There is no reason why Canada cannot adopt a similar strategy on an even more ambitious scale. For example, the federal government could invest in creating regulated sandboxes across the country, linked to universities. Each province would receive funding for sandboxes that allow the testing of AI technologies that are relevant to a specified economic strength. For example, while Nova Scotia could receive funding for testing AI innovations in fisheries, Saskatchewan could receive funding for AI applications related to mining, and Manitoba could receive funding for AI technologies related to farming. The regulated sandboxes could become a vital ecosystem where university research collaborates with local private sector innovation.

The production of more synthetic data is another area where Statistics Canada could make significant contributions to the AI ecosystem. As discussed by Sen et al. (2026), synthetic datasets resemble real-world health data in terms of statistical properties but are artificially generated, which enables researchers to perform meaningful analyses with a significantly reduced risk of privacy violations. With respect to the use of AI technologies in public healthcare, for example, synthetic data has shown promise in advancing importance. Specifically, synthetic data has been used to train AI models in areas such as disease prediction and clinical decision support (Giuffrè and Shung 2023; Gonzales, Guruswamy and Smith 2023).

Sen et al. (2026) further point out that the US provides clearer regulatory frameworks and more extensive access to anonymized and synthetic health datasets, enabling startups to develop, test, and refine AI-driven healthcare solutions more efficiently (also noted by Gonzales, Guruswamy and Smith 2023). Synthetic data holds significant

44 See: https://www.statcan.gc.ca/en/data-science/network/privacy-enhancing-techniques for further discussions.

45 See: https://www.gov.uk/government/calls-for-evidence/ai-growth-lab/ai-growth-lab.

potential to advance open data and open science initiatives in Canada while also supporting AI leadership and innovation (Arora 2023; Hu et al. 2023). Still, many existing datasets – particularly in sectors such as healthcare and finance – contain sensitive personal information, which restricts their availability due to privacy and confidentiality concerns (Public Health Agency of Canada 2022). By generating high-quality synthetic datasets that preserve the statistical properties of real data while protecting individual privacy, Canada can expand data accessibility for researchers, policymakers, and AI entrepreneurs.

Open science depends on the ability to validate and replicate empirical findings. Synthetic health data enable the sharing of datasets that closely resemble real-world distributions, thereby supporting reproducibility while remaining compliant with ethical and legal standards (Hu et al. 2023). Synthetic data can also facilitate stronger collaboration among public institutions, private organizations, and academic researchers by serving as a safe proxy for real-world data, enabling cross-sector innovation (Sharma 2024). Beyond research, synthetic datasets play a valuable role in education and workforce development by allowing students and professionals to engage with realistic data in hands-on training environments without risking exposure of sensitive personal information.

In summary, the key for Canada to be a global AI innovator is enhanced access to data for startups and other firms. For this to occur:

(1) Legal uncertainty over accessing online training data must be removed through appropriate amendments to privacy legislation;

(2) Public funding should be available to stimulate the creation of synthetic datasets;

(3) Private sector stakeholders should be enabled and encouraged to access confidential Statistics Canada and health data on a pilot basis, through appropriate PETs such as federated learning methods; and

(4) The federal government should invest in the creation of regulated sandboxes.

Except for China, other countries have not implemented strong policies to allow widespread data access, and this is where Canada can leverage its existing processes, which would allow researchers to employ confidential data. Enabling private sector access to confidential Statistics Canada data and creating regulated sandboxes as a mechanism to coordinate innovation between universities and the private sector could result in critical advantages to Canadian AI firms, as well as perhaps giving them an incentive to remain in Canada. Indeed, Canada can take a significant leap forward through enhanced data access to private sector organizations and by using PETs to protect individual privacy. Given its incredibly diverse population, Canadian data have the potential to reveal insights that might be relevant across the world.

Conclusion

Canada stands at an important moment in shaping its future as an AI global leader. While recent federal initiatives have made meaningful progress in strengthening research capacity, computing infrastructure, and talent development, these efforts will remain incomplete without a coordinated strategy to strengthen the country's data supply chains. AI innovation is fundamentally dependent on access to large, high-quality datasets, and the absence of clear, standardized, and legally supported data-sharing frameworks risks constraining the growth of Canadian AI firms and weakening international competitiveness.

Canada can differentiate itself internationally by building a governance model that balances expanded data access with rigorous protections for privacy and encourages public trust. Achieving this requires modernizing privacy legislation to explicitly recognize reidentification risks and incorporating social-benefit and cost-benefit tests into data-sharing decisions. At the same time, the strategic use of privacy-enhancing technologies – including DP, federated learning, and synthetic data – can significantly reduce the risk of harm to individuals

while preserving the analytical value of data for AI development.

In this respect, allowing carefully regulated private sector access to confidential Statistics Canada data, more convenient access to academic researchers, expanding the production of high-quality synthetic datasets, and establishing supervised regulatory sandboxes linked to regional strengths would substantially enhance Canada's AI ecosystem. These measures would support startups, reduce barriers to innovation, and foster collaboration among academia, government, and industry, while reinforcing public confidence through transparent governance mechanisms.

While some of the recommendations in this study may seem bold, they are consistent with the innovative data-sharing approach that the federal government is moving forward with through the introduction of Bill S-5, the *Connected Care for Canadians Act*. The bill is intended to enable Canadians to access their health data by mandating common standards for all information technology companies providing digital health services in Canada, as well as in the adoption of secure information exchange across various systems.[46]

Ultimately, the path forward lies not in choosing between innovation and privacy, but in designing institutions and policies that allow both to advance in tandem. By embedding trust, transparency, and social benefit at the core of its AI data strategy, Canada has the opportunity to create a globally influential model that drives inclusive economic growth, supports ethical AI development, and positions the country as a leader in responsible digital governance.

46 See: https://www.canada.ca/en/health-canada/news/2026/02/the-government-of-canada-introduces-legislation-to-build-a-more-connected-health-care-system.html for further details.

REFERENCES

Abowd, J. M., et al. 2025. "A Simulated Reconstruction and Reidentification Attack on the 2010 U.S. Census." *Harvard Data Science Review* 7(3). https://hdsr.mitpress.mit.edu/pub/ntchx9im/release/4.

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. 2022. "Too Much Data: Prices and inefficiencies in Data Markets." *American Economic Journal: Microeconomics* 14(4): 218–256. https://doi.org/10.1257/mic.20200200.

Arora, A. 2023. "Synthetic Data: The Future of Open-Access Health-care Datasets?" *The Lancet* 401(10381): 997. https://doi.org/10.1016/S0140-6736(23)00324-0.

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., and Weyl, E. G. 2018. "Should we treat data as labor? Moving beyond 'free.'" *AEA Papers and Proceedings* 108: 38–42. https://doi.org/10.1257/pandp.20181003.

Bernier, C. 2021. "Governance for Innovation and Privacy: The Promise of Data Trusts and Regulatory Sandboxes." Centre for International Governance Innovation (CIGI). https://www.cigionline.org/articles/governance-innovation-and-privacy-promise-data-trusts-and-regulatory-sandboxes/.

Council of Canadian Academies (CCA). 2023. "Connecting the Dots: The Expert Panel on Health Data Sharing." https://www.cca-reports.ca/reports/health-data-sharing-in-canada/.

Council of Canadian Academies. Council of Canadian Academies. 2015. "Accessing Health and Health-Related Data in Canada: The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation." https://cca-reports.ca/wp-content/uploads/2018/10/healthdatafullreporten.pdf.

Culnane, C., Rubinstein, B. I. P., and Teague, V. 2017. "Health Data in an Open World." *arXiv*. https://doi.org/10.48550/arXiv.1712.05627.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis. *Journal of Privacy and Confidentiality*, 7(3): 17–51. https://doi.org/10.29012/jpc.v7i3.405.

Girot, C., Shmerling Magazanik, L., and Sutherland, E. 2025. "We have a lot of valuable health data. Why is it so hard to use?" OECD Blog. https://www.oecd.org/en/blogs/2025/09/we-have-a-lot-of-valuable-health-data-why-is-it-so-hard-to-use.html.

Giuffrè, M., and Shung, D. L. 2023. "Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy." *npj Digital Medicine* 6(186). https://doi.org/10.1038/s41746-023-00927-3.

Gonzales, A., Guruswamy, G., and Smith, S. R. 2023. "Synthetic data in health care: A narrative review." *PLOS Digital Health* 2(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082.

Hu, B., Basri, M. A., Abdullah, A. Y., Tsao, S.-F., Butt, Z. A., and Chen, H. 2023. "Evaluation methods for synthetic data in pursuit of open data. *Proceedings of CVIS 2023* 9. https://openjournals.uwaterloo.ca/index.php/vsl/article/view/5860.

Jones, C. I., and Tonetti, C. 2020. "Nonrivalry and the Economics of Data." *American Economic Review*. 110(9): 2819–2858. https://doi.org/10.1257/aer.20191330.

Kashdaran, A. 2024. "Québec's New Data Portability Law: Key Features You Must Know." Privacy and Data Protection Bulletin. McMillan LLP. https://mcmillan.ca/insights/quebecs-new-data-portability-law-key-features-you-must-know/.

Lanier, J., and Weyl, E. G. 2018. "A Blueprint for a Better Digital Society." *Harvard Business Review*. https://hbr.org/2018/09/a-blueprint-for-a-better-digital-society.

Lehrer, S. F., and Xie, T. 2026. "What Lessons Should Canada Take on the Design of Public Data Exchanges?" *Canadian Public Policy* (Forthcoming).

Lu, Y., Kamath, G., and Yu, Y. 2023. "Exploring the Limits of Model-targeted Indiscriminate Data Poisoning Attacks." In *International Conference on Machine Learning*, pp. 22856-22879. PMLR.

Nair, A., and Inverarity, C. 2022. "What is Federated Learning?" The Open Data Institute. https://theodi.org/insights/explainers/what-is-federated-learning/.

Narayanan, A., and Shmatikov, V. 2006. "How to Break Anonymity of the Netflix Prize Dataset." *arXiv*. https://arxiv.org/abs/cs/0610105.

OECD. 2022. "Health Data Governance for the Digital Age." Paris: Organisation for Economic Co-operation and Development. https://www.oecd.org/en/publications/health-data-governance-for-the-digital-age_68b60796-en.

Public Health Agency of Canada. 2022. "Pan-Canadian Health Data Strategy Expert Advisory Group Report 3: Toward a World-class Health Data System." Government of Canada. https://www.canada.ca/en/public-health/corporate/mandate/about-agency/external-advisory-bodies/list/pan-canadian-health-data-strategy-reports-summaries.html.

Report of the House of Commons Standing Committee on Access to Information, Privacy and Ethics. 2022. "Collection and Use of Mobility Data by the Government of Canada and Related Issues." https://www.ourcommons.ca/Content/Committee/441/ETHI/Reports/RP11736929/ethirp04/ethirp04-e.pdf.

Savona, M. 2020. "Governance Models for Redistribution of Data Value." *VoxEU*. https://cepr.org/voxeu/columns/governance-models-redistribution-data-value

Scassa, T. 2022a. "Anonymization and De-identification in Bill C-27." https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=356.

_____. 2022b. "Data Sharing for Public Good: Does Bill C-27 Reflect Lessons Learned from Past Public Outcry?" https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=357.

Schwanen, D. 2023. "Getting Personal: The Promise and Potential Missteps of Canada's New Privacy Legislation." E-Brief 349. Toronto: C.D. Howe Institute. https://cdhowe.org/wp-content/uploads/2024/12/E-Brief_349-revised.pdf.

Schull, M., Paprica, P. A., Victor, J. C., and Saskin, R. 2017. "Institute for Clinical Evaluative Sciences (ICES) Exploratory Data & Analytic Services Private Sector Pilot Project." *International Journal of Population Data Science* 1(1): 88. https://doi.org/10.23889/ijpds.v1i1.88.

Sen, A. 2022. "Are Data Markets a Solution to Big Tech Market Power? A Competitive Analysis." *Journal of Government and Economics* 7: 100052. https://doaj.org/article/70efb5191c534416b2d030cb62a56414.

Sen, A., Chen, H., Grossman, M. R., and Tsao, S.-F. 2025. "A Welfare Test for Sharing Health Data." *Review of Income and Wealth*. 71(1): e12717. https://doi.org/10.1111/roiw.12717.

Sen, A., Tsao, S.-F., Chen, H., Meyer, S., Wheelans, C., and Abdulkarim, S. 2026. "Developing an Efficient Governance Framework for Synthetic Health Data for Canada." *Canadian Public Policy* (Forthcoming).

Sharma, K. 2024. *Evaluating Synthetic Data as a Proxy for Real Clinical Data in Machine Learning Models: A Comparative Study on Postpartum Hemorrhage Prediction*. Master's thesis, University of Waterloo. https://uwspace.uwaterloo.ca/items/287b5db6-73ab-4e50-aa98-49a4083a7f5c.

Siva Kumar, R.S., Penney, J., Schneier, B., and Albert, K. 2020. "Legal Risks of Adversarial Machine Learning Research." International Conference on Machine Learning (ICML) 2020 Workshop on Law & Machine Learning. July 3. https://ssrn.com/abstract=3642779.

Stanley, F., and Meslin, E. 2007. "Australia Needs a Better System for Health Care Evaluation." *Medical Journal of Australia* 186(5). https://www.mja.com.au/journal/2007/186/5/australia-needs-better-system-health-care-evaluation.

Statistics Denmark. 2014. "The Danish System for Access to Microdata." https://www.dst.dk/Site/Dst/SingleFiles/GetArchiveFile.aspx?fi=5452354440&fo=0&ext=israel2016.

Thalow, I., and Nielsen, C. 2015. "Access to Statistical Data for Scientific Purposes: New Nordic Model for Researchers' Joint Access to Data from the Nordic Statistical Institutions." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. https://unece.org/sites/default/files/datastore/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_4_Session_4_-_Denmark__Thaulow___Nielsen_.pdf.

NOTES:

NOTES:

Recent C.D. Howe Institute Publications

February 2026	Jones, David, and Tasnim Fariha. *Nation-Building in Practice: A Framework for Prioritizing Major Projects in Canada Canada's.* C.D. Howe Institute Commentary 705.
February 2026	Ambachtsheer, Keith, and Alex Mazer. *Spreading the Benefits: A Targeted Tax Credit Is Needed to Expand Retirement Plan Coverage in Canada's Private Sector.* C.D. Howe Institute Commentary 704.
January 2026	Yousif, James. *Accepting Asylum Claims Without a Hearing: A Critique of IRB's "File Review" Policy.* C.D. Howe Institute Commentary 703.
January 2026	MacKenzie, Peter, and Mark Zelmer. *The Window Is Closing: How Canada Can Shape the Future of Stablecoins and Digital Payments.* C.D. Howe Institute Commentary 702.
January 2026	Goulding, A.J. "Beyond "Energy Superpower": Market-Friendly Planning for Canada's Power Sector." C.D. Howe Institute E-Brief.
January 2026	Stokke, Anna. "Getting Math Instruction Right: Strategies for Improving Achievement in Canada." C.D. Howe Institute E-Brief.
December 2025	Fierlbeck, Katherine, and Rosalie Wyonch. *Disconnected: Inside Canada's Patchwork of Virtual Care*. C.D. Howe Institute Commentary 701.
December 2025	Zhang, Tingting. *The Quebec Primary Care Conundrum: Good Intentions, Persistent Problems*. C.D. Howe Institute Commentary 700.
Décembre 2025	Zhang, Tingting. *Le casse-tête des soins de bonnes intentions, des problèmes persistants*. Institut C.D. Howe commentaire 700.
December 2025	Robson, William B.P., and Mawakina Bafale. *Canada's Investment Crisis: Shrinking Capital Undermines Competitiveness and Wages*. C.D. Howe Institute Commentary 699.
November 2025	Kronick, Jeremy, Steve Ambler, and Thorsten Koeppl. "Mandating the Impossible? Assessing Canada's Electric Vehicle Mandate for 2026 and Beyond." C.D. Howe Institute E-Brief.
November 2025	Livingston, Brian. "Mandating the Impossible? Assessing Canada's Electric Vehicle Mandate for 2026 and Beyond." C.D. Howe Institute Working Paper.
November 2025	Robson, William B.P. "Fiscal Fantasies: Four Incredible Projections in the November 2025 Federal Budget." C.D. Howe Institute Verbatim.

Support the Institute

For more information on supporting the C.D. Howe Institute's vital policy work, through charitable giving or membership, please go to www.cdhowe.org or call 416-865-1904. Learn more about the Institute's activities and how to make a donation at the same time. You will receive a tax receipt for your gift.

A Reputation for Independent, Nonpartisan Research

The C.D. Howe Institute's reputation for independent, reasoned and relevant public policy research of the highest quality is its chief asset, and underpins the credibility and effectiveness of its work. Independence and nonpartisanship are core Institute values that inform its approach to research, guide the actions of its professional staff and limit the types of financial contributions that the Institute will accept.

For our full Independence and Nonpartisanship Policy go to www.cdhowe.org.

C.D. HOWE
INSTITUTE

110 Yonge Street, Suite 800,
Toronto, Ontario
M5C 1T4

Canadian Publication Mail Sales
Product Agreement #40008848