SENG 360 - Security Engineering
Database Security - Inference Control

Jens Weber

Fall 2021

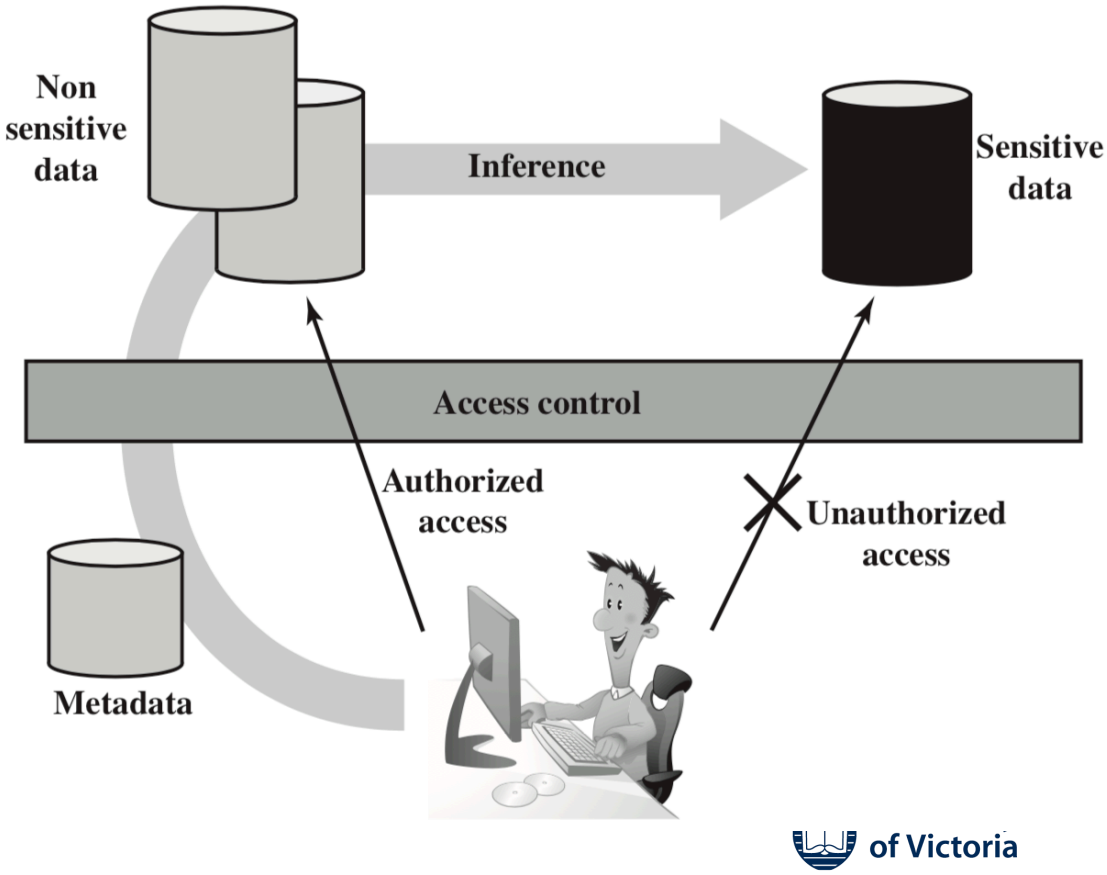
Chapter 11, Textbook

University of Victoria

Inference

Inference is the process of performing authorized queries for deducing unauthorized information

The inference problem arises when the combination of non-sensitive data may result in sensitive data.

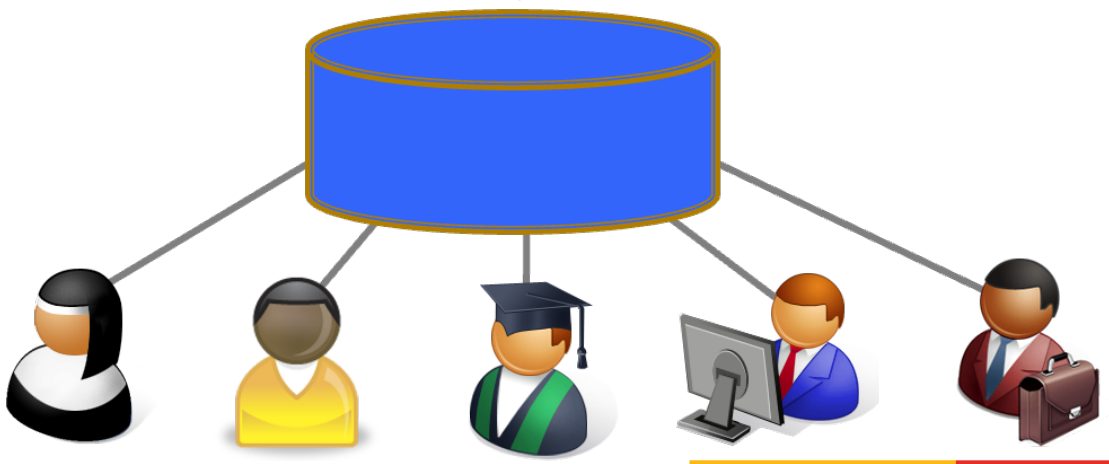


of Victoria

General Setting

Medical data
Query logs
Social network data
…

Data mining
Statistical queries



University of Victoria

Anonymized data - is there such a thing?

"Anonymized data" is one of those holy grails, like "healthy ice-cream" or "selectively breakable crypto".

– CORY DOCTOROW

University of Victoria

Learning Objectives

At the end of this class you will be able to

- Define what an inference attack is
- Describe different approaches to inference control
- Describe the idea off differential privacy

University of Victoria

"Anonymized" data really isn't—and here's why not

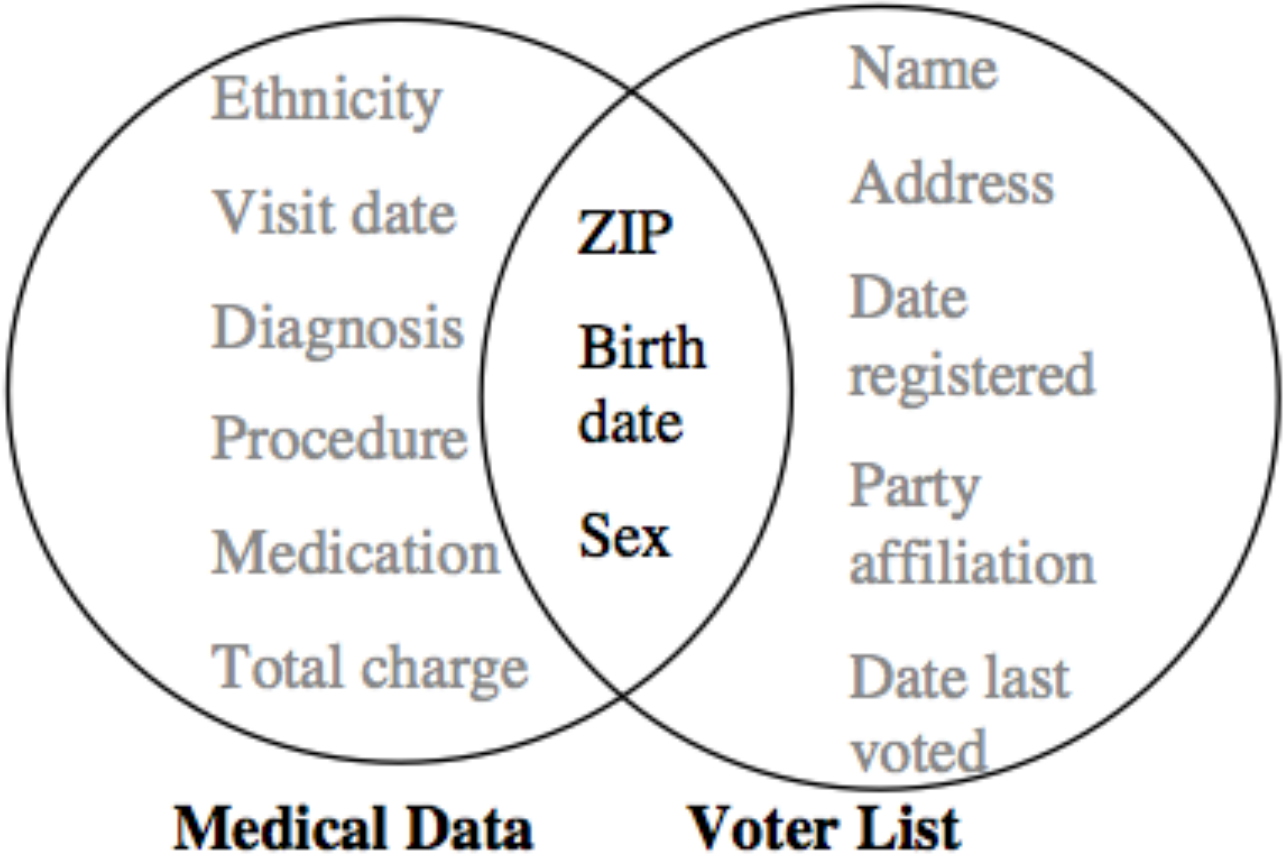
Companies continue to store and sometimes release vast databases of " …

by **Nate Anderson** - Sep 8, 2009 3:25am PST

The Massachusetts Group Insurance Commission had a bright idea back in the mid-1990s—it decided to release "anonymized" data on state employees that showed every single hospital visit. The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number. But a graduate student in computer science saw a chance to make a point about the limits of anonymization.

Latanya Sweeney requested a copy of the data and went to work on her "reidentification" quest. It didn't prove difficult. Law professor Paul Ohm describes Sweeney's work:

At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.



Sweeny (2000) 87% of the population of the United States can be uniquely identified by gender, date of birth, and 5-digit zip code

Simple IC: Query Size Control

query must involve at least target k

University of Victoria

Simple IC: Query Size Control

Let's assume there is only one male Full Professor in the department

query average salary of all male professors in the dept (more than k)

query average salary of all full profs in the dept (more than k)

University of Victoria

Simple IC: Query Size Control

query must involve at least target k

query must at most target N-k records
(N total number of records)

University of Victoria

Trackers

The ‘professor’ queries are example for a **tracker** attack

- Attacker partitions knowledge in multiple queries (that do not violate the size restriction) and than combines results
- Decompose Q (violating size restriction) into two parts C= C1 · C2, such that C1 and T = (C1 · ~C2) satisfy size restriction

University of Victoria

Inference Control at Query Time

Idea: keep track of past queries for given user and determine whether new query should be allowed

However, not allowing a new query may also reveal sensitive info.

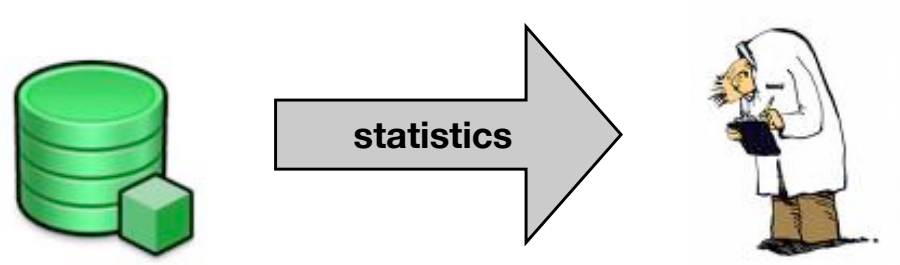
Example: *Individual salaries are sensitive*

User queries: Select SUM(salary) from Salaries; 150,000

Use queries: Select COUNT(salary) from Salaries; 3

User queries: Select MAX(salary) from Salaries; 50,000

Query Overlap Control



Idea: keep track of all past queries of a given user

New query C is allowed only if overlap
with past queries D is less than ***r*** records: $|X(C) \cap X(D)| \leq r$

Perturbation

Query restriction can be costly and difficult to implement

Alternative: modify (perturbe) the data

Two options:

1. Perturbe the stored data (Data Perturbation)
2. Perturbe the output of queries (Output Perturbation)

Note: Perturbation should preserve statitistical results as much as possible

University of Victoria

Cell Suppression

Example: Average grades of Major/Minor s… minimum query size 3

Major:	Biology	Physics	Chemist…y	Geology
Minor:				
Biology	-	blanked	17	11
Physics	7	-	32	18
Chemistry	33	blanked	-	blanked
Geology	9	13	6	-

can be used to reconstruct

if known

if known

suppress (because only 2 students in class)

Major:	Biolog…		…hemistry	Geology
Minor:				
Biology	-	blanked	17	blanked
Physics	7	-	32	18
Chemistry	33	blanked	-	blanked
Geology	9	13	6	-

can be used to reconstruct

Blanking a cell in a DB with m tuples means blanking 2m-1 other cells

Swapping

Swap a sufficiently large number of attributes such that certain desired statistical properties are preserved

	D			D′		
Record	Sex	Major	GP	Sex	Major	GP
1	Female	Bio	4.0	Male	Bio	4.0
2	Female	CS	3.0	Male	CS	3.0
3	Female	EE	3.0	Male	EE	3.0
4	Female	Psy	4.0	Male	Psy	4.0
5	Male	Bio	3.0	Female	Bio	3.0
6	Male	CS	4.0	Female	CS	4.0
7	Male	EE	4.0	Female	EE	4.0
8	Male	Psy	3.0	Female	Psy	3.0

Example: Count(Female · CS), Count(Male · 4.0)

University of Victoria

k-anonymity

Sweeney (2002)

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053			
2	13068			
3	13068			
4	13053			
5	14853			
6	14853			
7	14850			
8	14850			
9	13053			
10	13053			
11	13068			
12	13068	35	American	Cancer

Background information attack

Homogeneity attack

Machanavajjhala (2006)

An improvement: **l -diversity**

Machanavajjhala (2006)

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

4-anonymous data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

3-diverse data

Further improvement: t-closeness

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

3-diverse data

Similarity attack

Li et al (2007)

Bob (47602, 22 years) has a stomach problem

University of Victoria

Output Perturbation

Technique 1: **Random-sample Query**

Execute user's query on a randomly selected sample of all records in the database

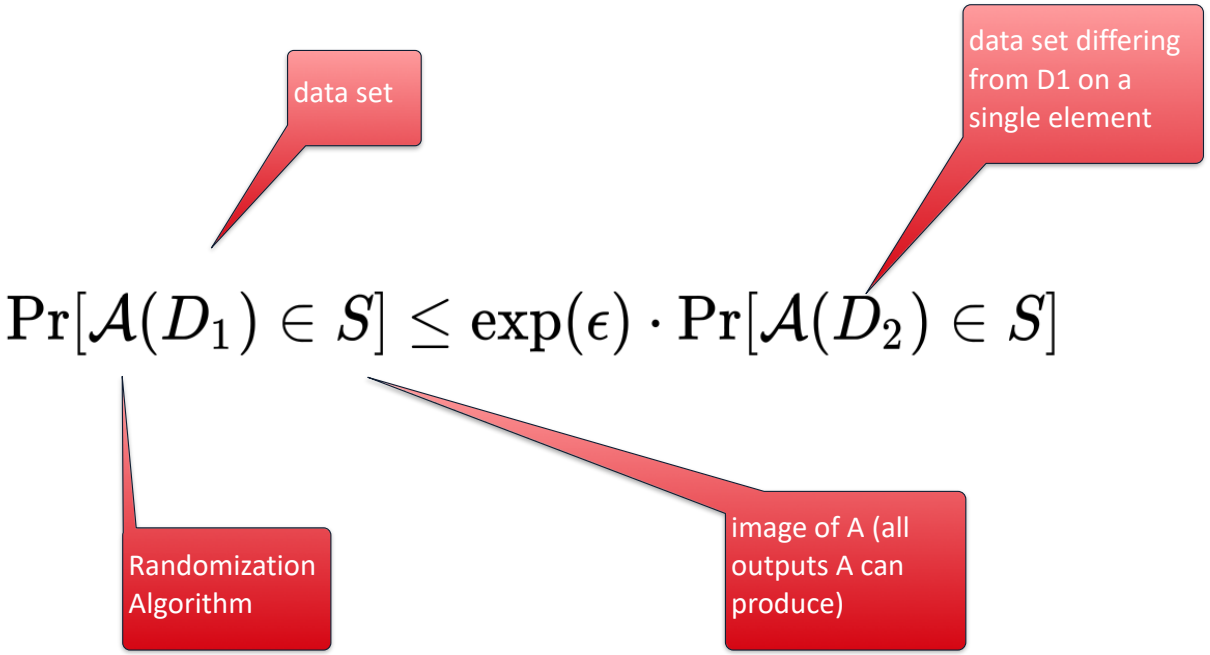
Works only for large datasets

Technique 2: **Add random "noise" to query answer**

University of Victoria

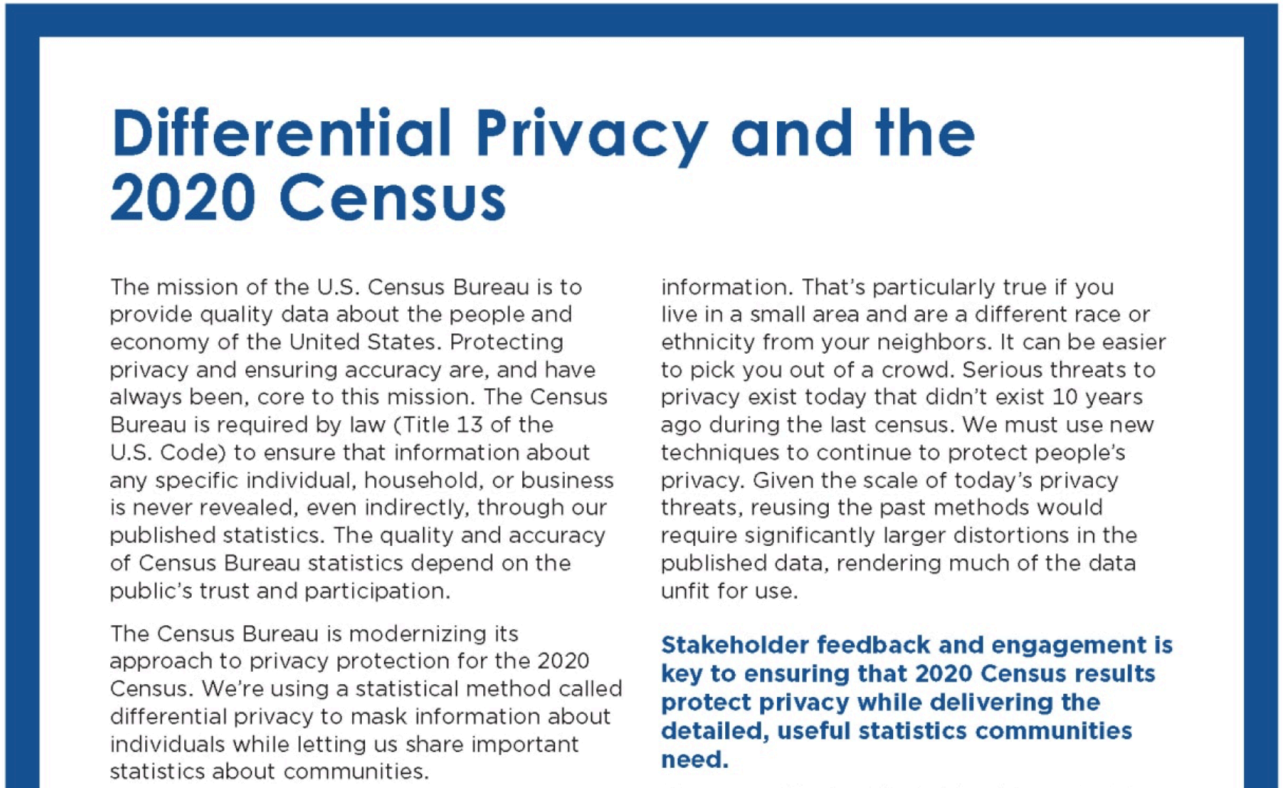
ϵ-Differential Privacy

Mathematical definition of privacy loss associated with statistical data release



Differential Privacy and the 2020 Census

JULY 22, 2021



Differential Privacy and the 2020 Census

The mission of the U.S. Census Bureau is to provide quality data about the people and economy of the United States. Protecting privacy and ensuring accuracy are, and have always been, core to this mission. The Census Bureau is required by law (Title 13 of the U.S. Code) to ensure that information about any specific individual, household, or business is never revealed, even indirectly, through our published statistics. The quality and accuracy of Census Bureau statistics depend on the public's trust and participation.

The Census Bureau is modernizing its approach to privacy protection for the 2020 Census. We're using a statistical method called differential privacy to mask information about individuals while letting us share important statistics about communities.

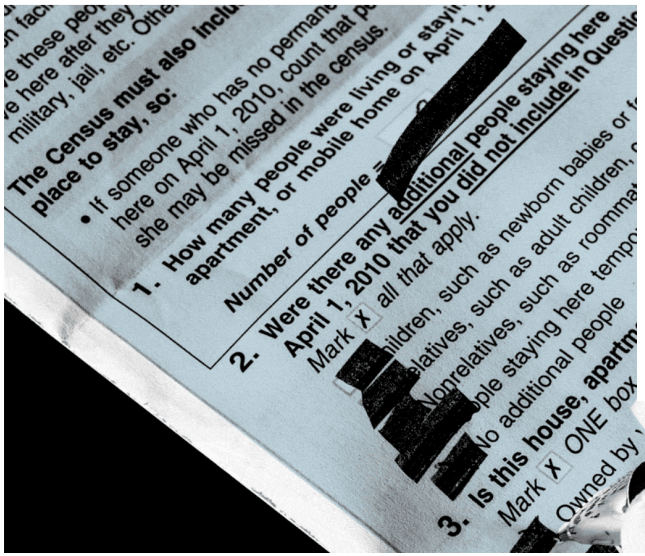
information. That's particularly true if you live in a small area and are a different race or ethnicity from your neighbors. It can be easier to pick you out of a crowd. Serious threats to privacy exist today that didn't exist 10 years ago during the last census. We must use new techniques to continue to protect people's privacy. Given the scale of today's privacy threats, reusing the past methods would require significantly larger distortions in the published data, rendering much of the data unfit for use.

Stakeholder feedback and engagement is key to ensuring that 2020 Census results protect privacy while delivering the detailed, useful statistics communities need.

future tense

States Are Suing the Census Bureau Over Its Attempts to Make Data More Private

BY PRIYANKA NANAYAKKARA AND JESSICA HULLMAN AUG 12, 2021 • 5:50 AM



Although 2020 Census statistics have yet to be fully released, the bureau is already facing pushback about the new approach. Alabama is suing the bureau (though the lawsuit was recently put on hold) on the grounds that it is violating its mandate to "report … accurate '[t]abulations of the population' " by releasing "inaccurate" statistics. Sixteen states are backing the lawsuit. The bureau, however, is also legally required not to publish individually identifiable data, putting it between a rock and a hard place. *Some* approach to protecting privacy is mandatory and differential privacy represents the strongest known defense against re-identification.

Frequently asked questions—Security and privacy

Expand all Collapse all

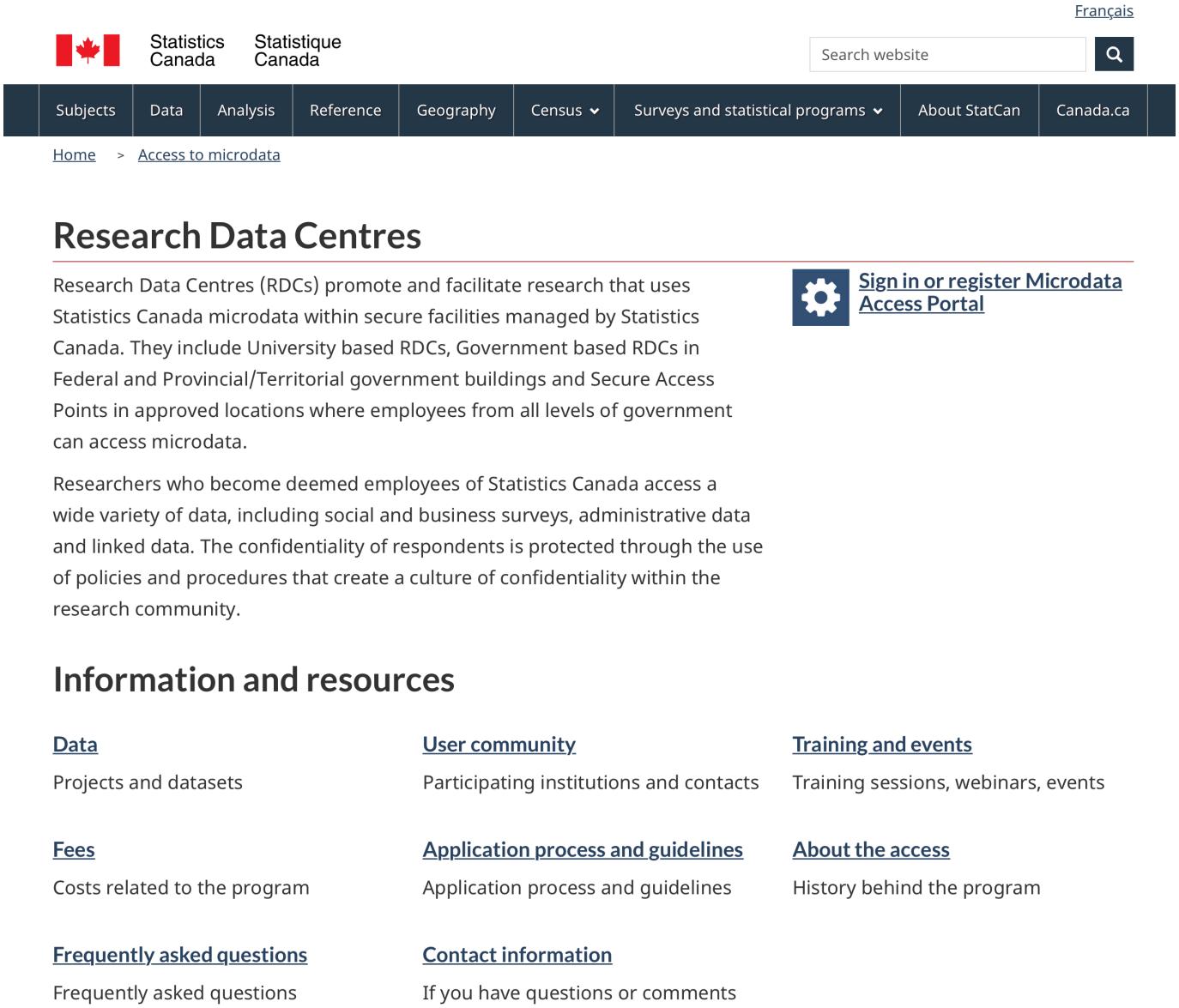
▼ 10. How does Statistics Canada ensure the confidentiality of the information it publishes?

Statistics Canada is bound by law to protect the identity of individuals in any data it publishes. Publications and electronic data releases are screened so that anonymity is assured. Names, addresses and telephone numbers are not part of the census database used for dissemination, and private contractors do not have access to confidential data.

Published census data go through a variety of automated and manual processes to determine whether the data need to be suppressed. This is done primarily to ensure that the identity and characteristics of respondents are not disclosed (referred to as confidentiality).

Confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data. Consequently, the agency does not publish data from geographic areas with a population below a certain threshold.

Combine IC and AC



Français

Statistics Canada Statistique Canada

Search website

Subjects | Data | Analysis | Reference | Geography | Census | Surveys and statistical programs | About StatCan | Canada.ca

Home > Access to microdata

Research Data Centres

Research Data Centres (RDCs) promote and facilitate research that uses Statistics Canada microdata within secure facilities managed by Statistics Canada. They include University based RDCs, Government based RDCs in Federal and Provincial/Territorial government buildings and Secure Access Points in approved locations where employees from all levels of government can access microdata.

Researchers who become deemed employees of Statistics Canada access a wide variety of data, including social and business surveys, administrative data and linked data. The confidentiality of respondents is protected through the use of policies and procedures that create a culture of confidentiality within the research community.

Sign in or register Microdata Access Portal

Information and resources

Data
Projects and datasets

User community
Participating institutions and contacts

Training and events
Training sessions, webinars, events

Fees
Costs related to the program

Application process and guidelines
Application process and guidelines

About the access
History behind the program

Frequently asked questions
Frequently asked questions

Contact information
If you have questions or comments

10	**11**	**12** 10-2p	**13** 9-4p	**14** 10-2p	**15**	16
17	**18**	**19** 10-2p	**20** 9-4p	**21** 10-2p	**22**	23
24	**25**	**26** 10-2p	**27** 9-4p	**28** 10-2p	**29**	30
31						

For any questions or concerns, please contact the RDC staff at rdc@uvic.ca or 250-853-3196.
You can also contact Dr. Herb Schuetze, Academic Director of the UVic branch of the BCIRDC, at 250-721-8541 or hschuetz@uvic.ca

Data Access

Researchers with an approved project can access RDC data. Applications can be made by individual researchers or by research teams led by a principal applicant. Please visit the Statistics Canada Microdata Access Portal to apply.

The Statistics Canada Data Liberation Initiative (DLI) provides direct access to Public Use Microdata Files (PUMFs). Applicants must demonstrate that their research cannot be conducted using PUMFs. UVic affiliates have access through the Library Data Services Dataverse.

Summary and Outlook

- Inference attacks disclose sensitive data by combining non-sensitive data sets
- Inference control (IC) important concern in statistical databases
- Techniques: query control and pertubation
- Trade-off between precision and privacy

- Next week: Access Control Models

University of Victoria

Questions?

University of Victoria