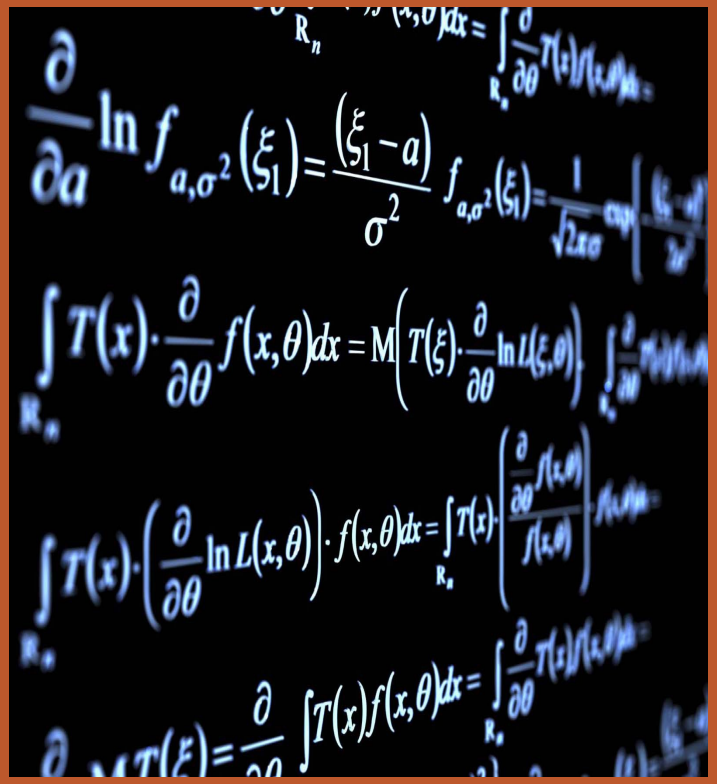


Discussion with MLIA

Implications to Montana of Differential Privacy for Census Bureau Data Dissemination

Mary Craigle, RIS Bureau Chief
MT Department of Commerce
April 22, 2020

High Level Overview of Differential Privacy (DP)

Highlights in the history of census privacy

1929: Census law made protection explicit

"No publication shall be made by the Census Office whereby the data furnished by any particular establishment or individual can be identified, nor shall the Director of the Census permit anyone other than the sworn employees to examine the individual reports."

1954: Title 13 retained 1929 language

1962: No sharing within government, immune from legal process

2002: Confidentiality requirements clarified by the "Confidential Information Protection and Statistical Efficiency Act" (CIPSEA) formally defined the meaning of identifiable data

DP is a step beyond what traditionally has been Disclosure Avoidance

The new disclosure rules were motivated by the threat of "database reconstruction" inferring individual-level data from tabular data.

Database reconstruction should not be confused with re-identification. To identify anyone's characteristics, one would have to match the reconstructed microdata to another source that provides the individual's information.

Since 1962, the Census Bureau has interpreted "any particular establishment or individual" to mean an individual whose identity can be determined. Now the Census Bureau is saying it cannot release data about individuals, **even if the identity of those individuals is unknown**, because they could be identified using information from another source.

The new interpretation asserts that it is **prohibited to reveal characteristics of an individual even if the identity of that individual is effectively concealed. This is because the microdata records could be linked to a commercial database to determine PII to determine race and ethnicity.**

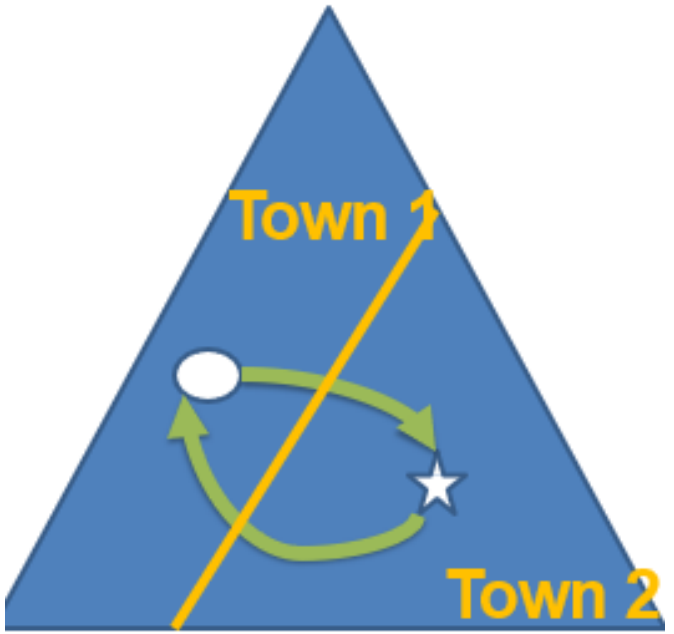
The protection system used in 2000 and 2010 relied on swapping households.

Some households were swapped with other households

- Swapped households had the same size.
- Swapping limited to within each state.

Disadvantages:

- Swap rate and details of swapping not disclosed.
- Privacy protection was not quantified.
- Impact on data quality not quantified



State "X"

The Disclosure Avoidance System Relies on Injecting Statistical Noise into the Data with Formal Privacy Rules

Advantages of noise injection using differential privacy:

- Privacy guarantees are *future-proof*
- Privacy guarantees are ***provable***
- Privacy guarantees are *public and explainable*
- Protects against database reconstruction attacks (tunable)

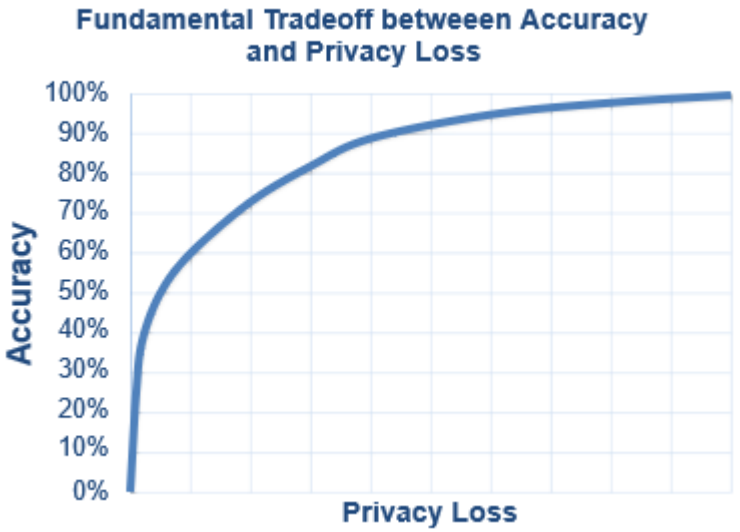
Disadvantages:

- Entire country must be processed at once for best accuracy
- Every use of the private data must be tallied in the *privacy-loss budget*

D{ is the formal privacy system the Census Bureau will be using on the 2020 Census Data and ALL other data including that prepared for other agencies (BLS, BEA, etc.)

DP features:

- Provable bounds on the maximum privacy loss
- Algorithms that allow policy makers to manage the trade-off between accuracy and privacy
- It's called "differential privacy" because it mathematically models the privacy "differential" that each person experiences from having their data included in the Census Bureau's data products compared to having their record deleted or replaced with an arbitrary record.



How DP would impact the data that GIS Professionals (and a plethora of others) use.

Major Impacts of DP to Montana Data are in three areas:

1. For data that is release, the smaller the geography size, the more distorted the data.

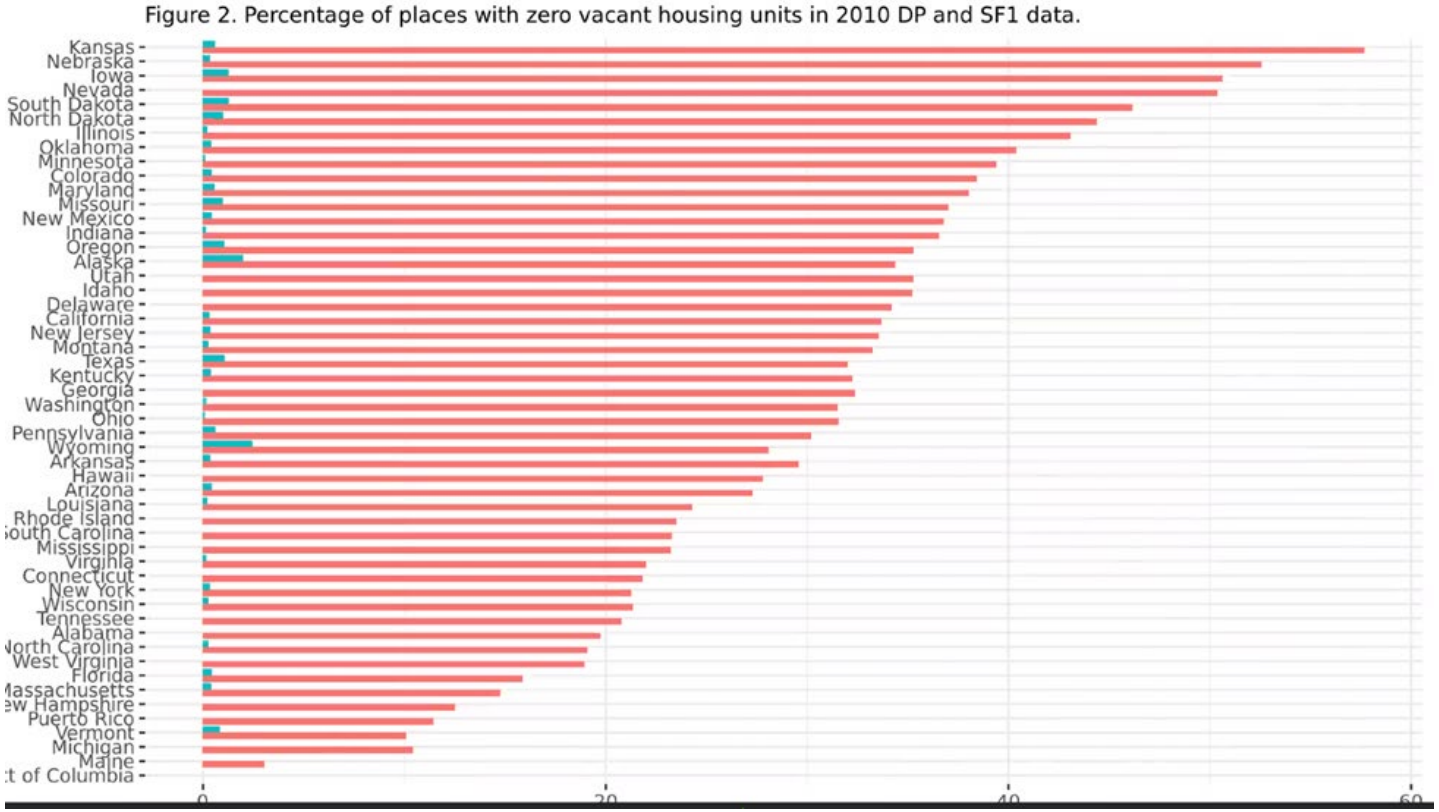
2. Because of the loss of accuracy, much of the data available in the past from all the various Census, BLS, BEA, etc. sources would be suppressed due to the lack of accuracy.

3. There is significant impact to all the many users including government that uses the data for funding, planning, and assessment, researchers who use the data for study, and businesses and non-profits who is the data for planning and marketing.

The ACS is among the most widely-used scientific data sources in the world. Google Scholar lists 55,000 references to the ACS, and on average a new paper using the data appears every 55 minutes.

Examples from the 2010 Demonstration Product where DP was applied

The 2010 Decennial Census data reported only **one community** in Montana that had no unoccupied housing units – Saddle Butte Census Designated Place (CDP) which is a which is in Hill County along the southeast border of Havre. With DP applied to this same dataset, approximately 120 of the 364 CDPS in Montana would have reported no vacant housing units.

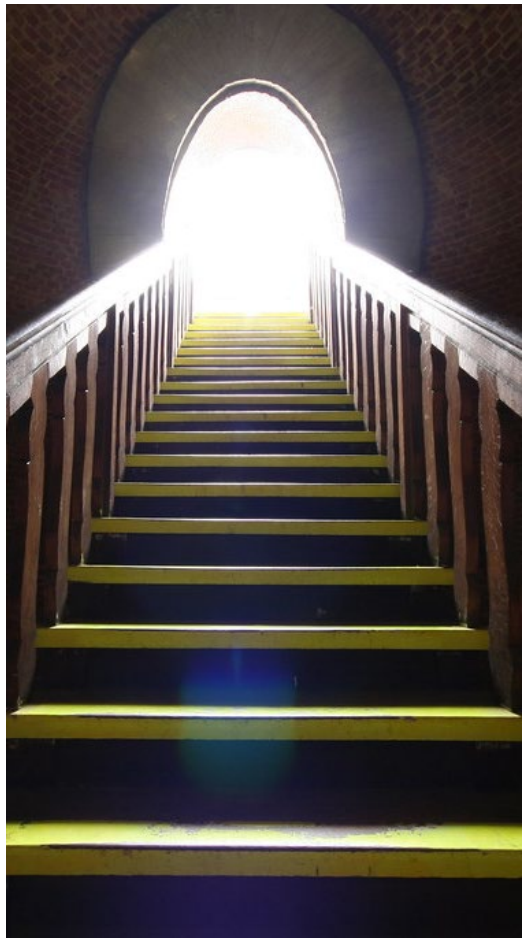


Gray bars are the number of CDPs with no vacant housing units from the 2010 Decennial Census data as published. Red bars are the CDPs with no vacancies when DP was applied to the data set.

Examples from the 20Examples from the 2010 Demonstration

As another example, the number of households with female householder, no husband present with own children under 18 years decreases from the 2010 reported of 87 to 67 households in Cut Bank, MT with DP applied.

For Crow Agency, the reported number changes from 2010 reported of 37 to 44 households. Again, distorting the data through the application of DP will have major impacts in childcare funding and planning in this community and across Montana.



Steps that have been taken thus far by folks from the Professional Community (including State Data Centers (SDC) Census Information Centers (CIC), Federal Statistical Co-op for Population Statistics (FSCPE) & Demographers and Researchers.