Risk assessment procedures for the 2020 U.S. Census

Steven Ruggles
Lara Cleveland
David Van Riper

Institute for Social Research and Data Innovation
University of Minnesota

UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality (1 December 2021)

Problem statement

- US Census Bureau justified its transition to a formally private disclosure avoidance system based on the results of a database reconstruction and re-identification experiment.

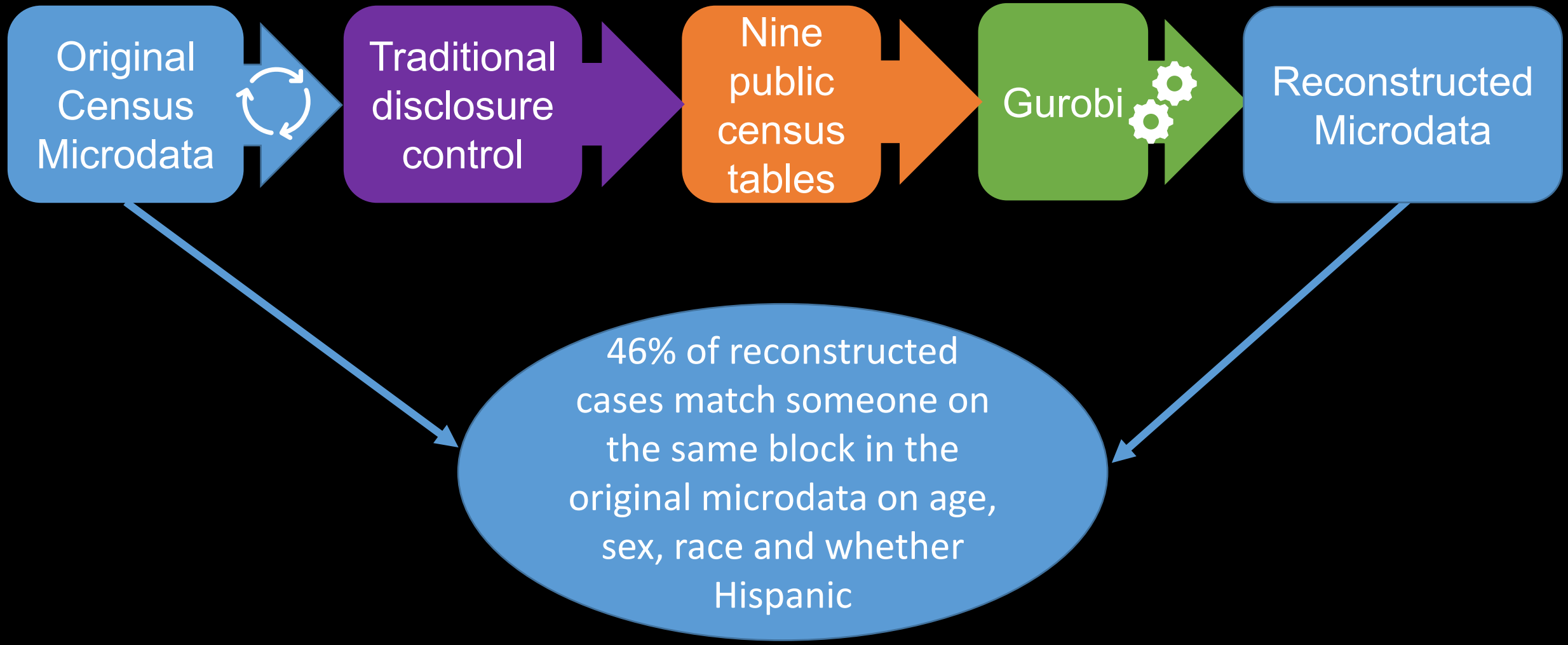
Problem statement

- US Census Bureau justified its transition to a formally private disclosure avoidance system based on the results of a database reconstruction and re-identification experiment.
- How does the Bureau's results compare with a similar (reconstruction) experiment based on chance?

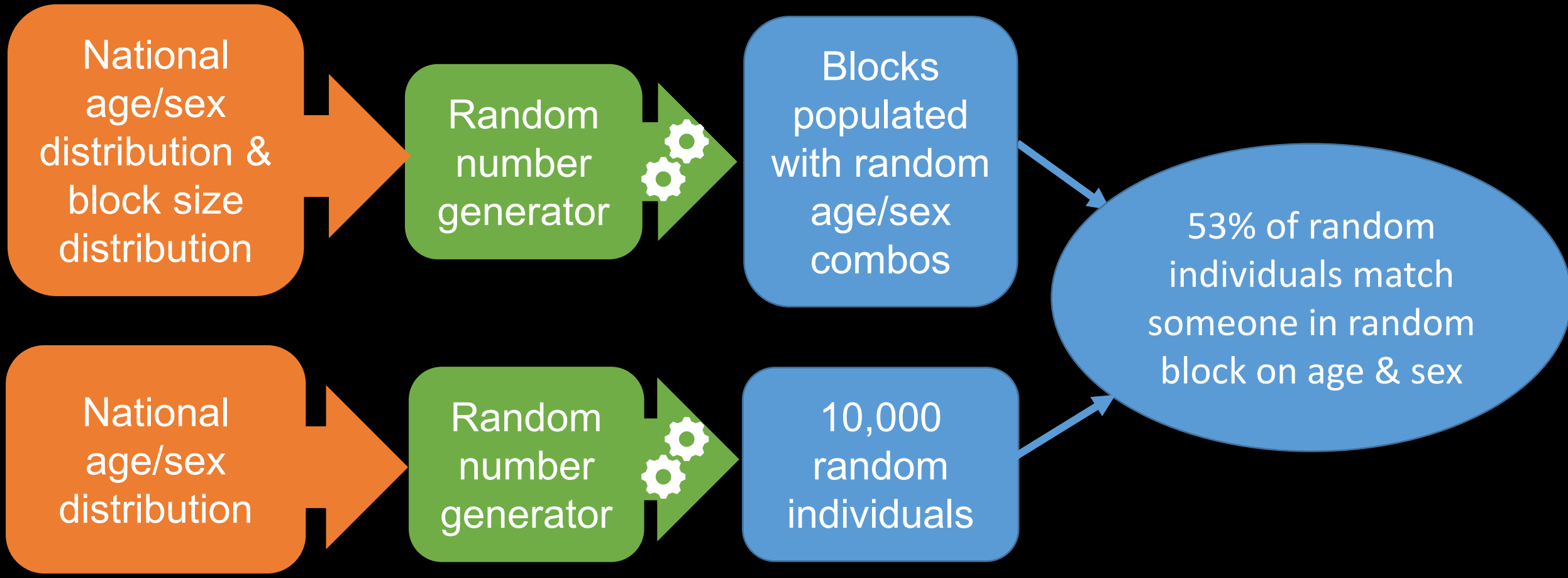
Reconstruction Results

- 46% of Bureau's reconstructed cases match someone in confidential data on census block, age, sex, race, Hispanic
- 41% of our random individuals match someone in random block on age, sex, race, Hispanic ethnicity

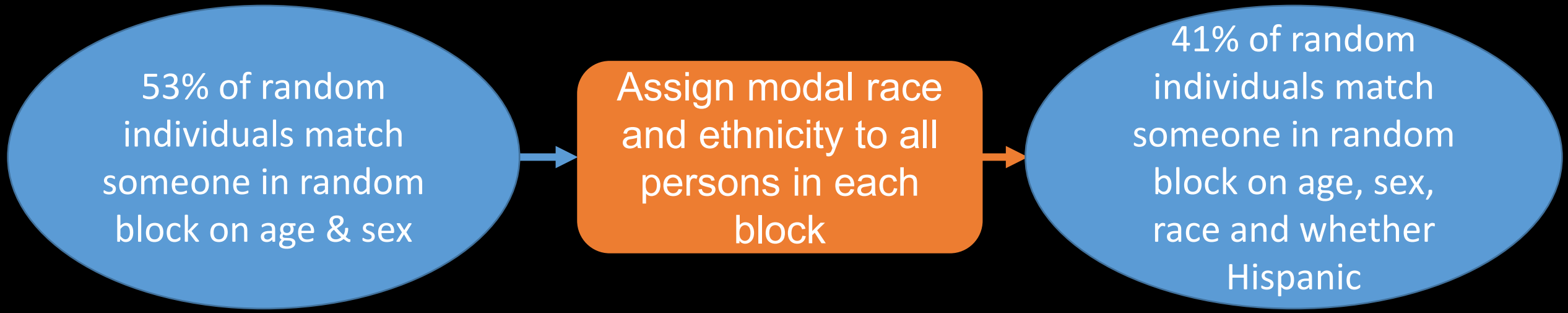
Census Bureau Database Reconstruction Experiment



How many matches would be expected by chance?



Adding Race and Ethnicity



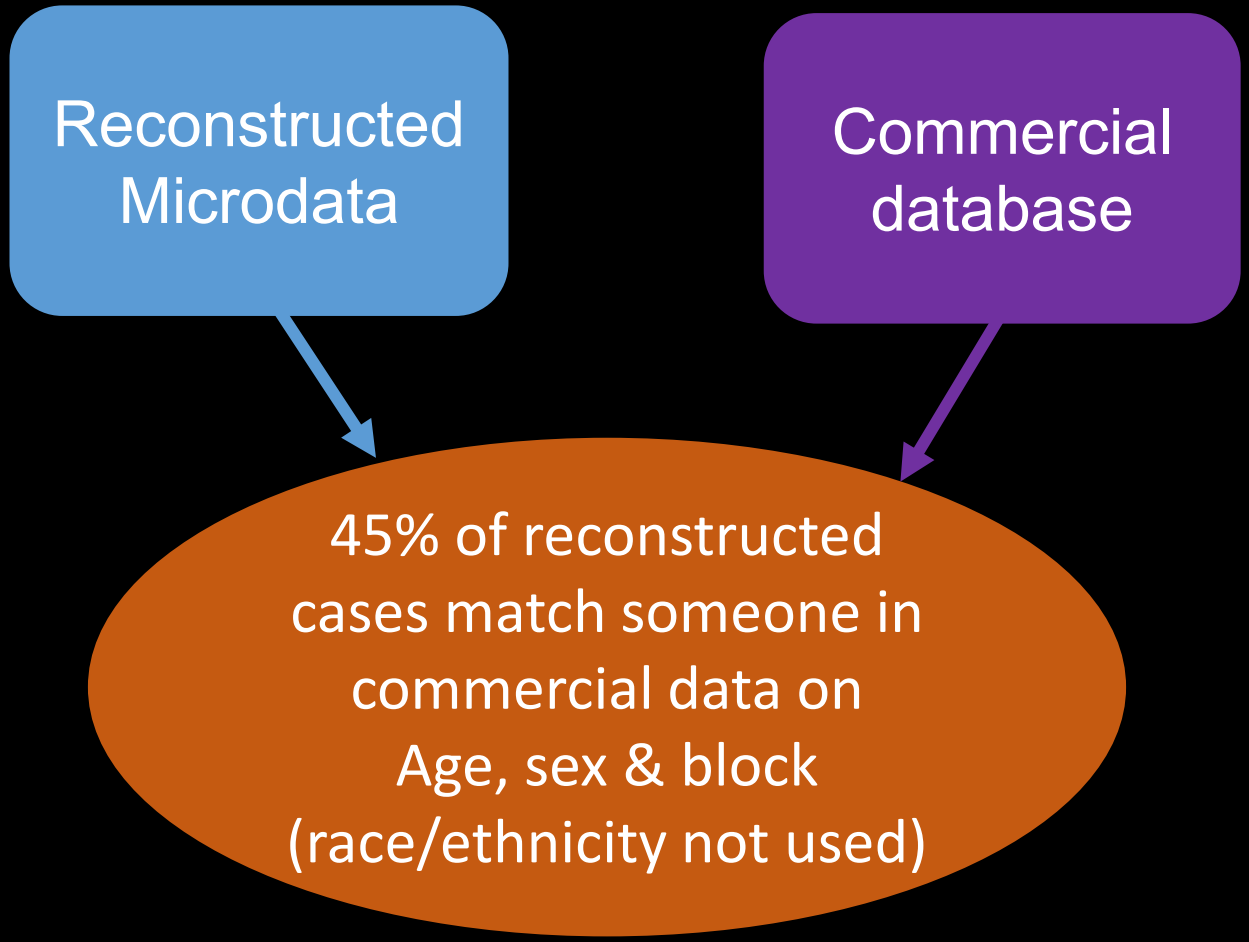
Database reconstruction experiment

- The Census Bureau “Reconstruction” matched on block, age, sex, and race/ethnicity in 46% of cases
- A random number generator combined with a simple assignment rule for race and ethnicity yields a comparable match rate of 41%
- Despite the Census Bureau’s massive investment of resources and computing power, their database reconstruction does not perform much better than a roll of the dice
- This is analogous to a clinical trial in which the treatment and the placebo produce virtually the same outcome.

Re-identification Results

- 45% of Bureau's reconstructed cases match someone in commercial data on census block, age, and sex
- 38% of names harvested from commercial database match a name on same block in confidential data
- 17% (38% of 45%) of Census respondents "re-identified"

Census Bureau rei-dentification experiment

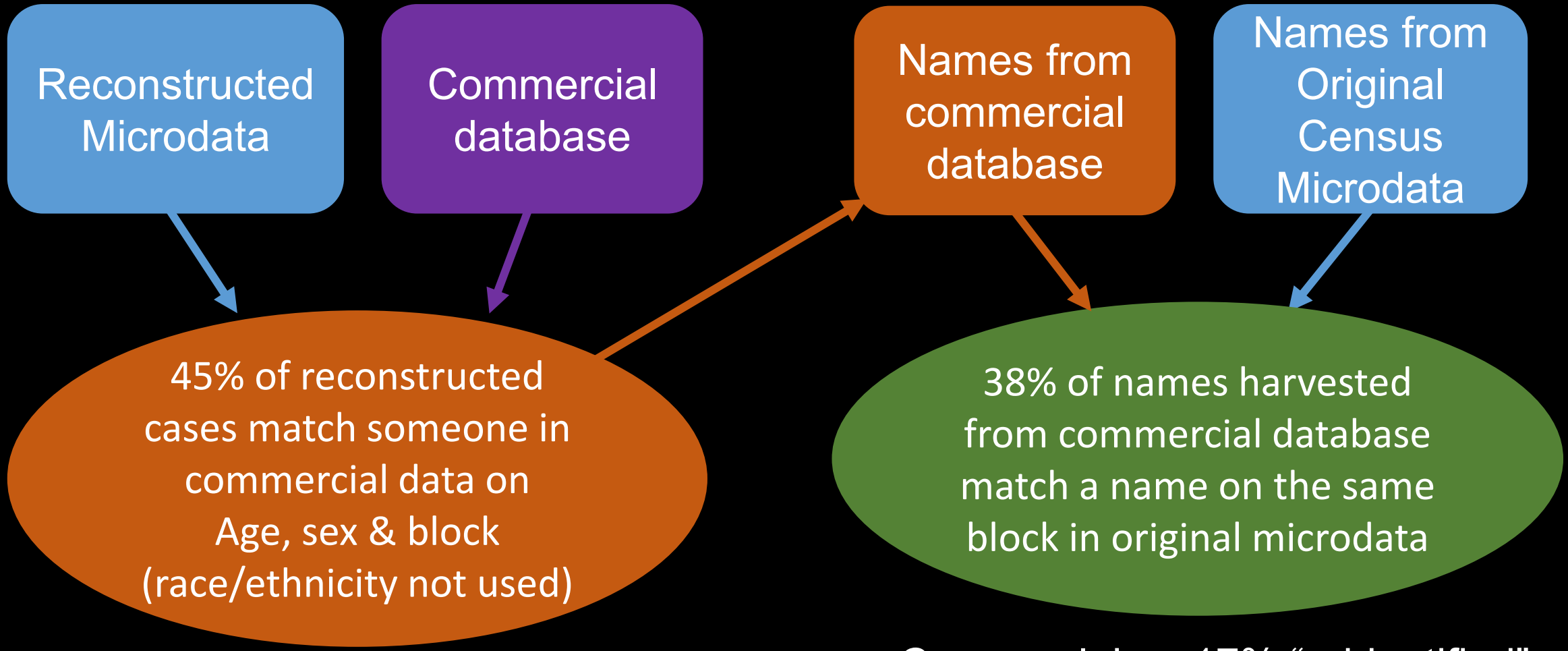


(One would expect 53% by chance)

The re-identification experiment

- Within each block the Census Bureau searched a commercial database for people who matched the age and sex of each case of their reconstructed database.
- In just 45% of cases was there at least one person in the commercial data who matched the age, sex and block number of a row of the hypothetical database.
- This is a lower match rate than the 53% one would expect by chance, probably because of the poor quality of the commercial data.

Census Bureau re-identification experiment



(One would expect 53% by chance)

Census claims 17% “reidentified” (38% of 45%)

The re-identification experiment

- The Census Bureau then harvested the names of the 45% of cases in the commercial database that matched on age and sex and searched for them in the original census data.
- They found that 38% of the 45% were actually present on the block.
- They claimed to have “reidentified” 16.85% of the population (38% of 45%).

Census Bureau re-identification experiment

Among the 45% of cases where
the reconstructed data match
someone in the commercial data
on age, sex, and block,
38% match a name on the same
block in original microdata

- There is no null comparison
- One would expect that people residing on a particular block in the commercial data would also be enumerated in the census.
- Is 38% high or low?

Evaluating the re-identification experiment

- The Census Bureau could easily evaluate the re-identification by matching the names of people randomly selected from the commercial database to persons in the 2010 census living on the same census block, without any reference to the Census Bureau's database reconstruction.
- If the 38% match rate on names for the reconstructed population is no higher than the match rate for a randomly selected subset of the commercial data, it would mean the database reconstruction had no effect on re-identification risk.
- Because they made no null comparison, the re-identifcation results cannot be interpreted.

Summary: Database reconstruction experiment

- “Correctly” guessed age, sex, race and Hispanic ethnicity for 46% of persons on each block
- Great majority of those correct guesses would be expected by chance
- An outside attacker would have no means of determining which “reconstructed” records were true

Summary: Re-identification experiment

- The re-identification experiment shows only that some people found in a commercial database are also listed on the same block in the census.
- It fails to show that using the published tabulations had any impact whatsoever on the match rate between the commercial database and the original enumeration.

Consequences of disclosure panic

The disclosure avoidance system introduces unacceptable levels of error for many applications of the census.

- *Science Advances*
- *Pop. Res. Policy Rev*
- *Socius*
- *PNAS*

SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census

Christopher T. Kenny[1], Shiro Kuriwaki[2], Cory McCartan[3], Evan T. R. Rosenman[4], Tyler Simko[1], Kosuke Imai[1,3]*

Census statistics play a key role in publ revealing individual information, many differential privacy, which add noise to census statistics must be postprocessed Bureau's latest disclosure avoidance sy electoral districts. We find that the DAS sy precincts, yielding unpredictable raci "One Person, One Vote" standard as curr race and ethnicity. Our findings underscor

INTRODUCTION

In preparation for the official release of U.S. Census Bureau has developed a di (DAS) to prevent Census responses fro individuals (*1*). The DAS is based on diff which adds a certain amount of random The Bureau has been required by law t information about Census participants implemented disclosure avoidance met their decision to incorporate differentia subsequent postprocessing steps in the 20 in the DAS, has been controversial. Som cerns about the potential negative impa policy and social science research, which data (*2*–*6*).

The U.S. decennial census serves as case study on the impact of differential the drawing of legislative districts, det federal funds for more than a hundred g are extensively analyzed by social scient and international organizations, inclu United Kingdom, and Australia, have a the adoption of differential privacy tech to its decennial census, the U.S. Census B ferential privacy as their "privacy definitio data release on commuting patterns (*12* sidering adopting a similar approach fo as the American Community Survey (*13*

It is a common misconception that a only involves injecting random noise (*14*

Population Research and Policy Review
https://doi.org/10.1007/s11113-021-09664-5

RESEARCH BRIEFS

Check for updates

Differential Privacy and the Accuracy of County-Level Net Migration Estimates

Richelle L. Winkler[1] · Jacl David Egan-Robertson[3]

Received: 5 March 2021 / Accepted

Abstract

Each decade since the 195 tion estimates by age, sex as starting and ending po thousands of times and wi Census 2020 should allow accuracy of new estimates (DP) disclosure avoidance brief estimates the impact migration estimates. Usin we construct a hypothetica pare them to published esti sis. Findings show that ba

Check for updates

SOCIUS

Original Article

ASA American Sociological Association

Differential Privacy in the 2020 Census Will Distort COVID-19 Rates

Socius: Sociological Research for a Dynamic World
Volume 7: 1–6
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2378023121994014
srd.sagepub.com
SAGE

Mathew E. Hauer[1]

Abstract

Scholars rely on accurate (COVID-19) pandemic, wit deaths at older ages. Popu subject to noise infusion on differential privacy. Us introduce substantial dist percent, hindering our abil with fewer than 1,000 pers U.S. Census Bureau shoul differential privacy's distor

Check for updates

How differential privacy will affect our understanding of health disparities in the United States

Alexis R. Santos-Lozada[a,1], Jeffrey T. Howard[b], and Ashton M. Verdery[c]

[a]Department of Human Development and Family Studies, The Pennsylvania State University, University Park, PA 16802; [b]Department of Public Health, University of Texas at San Antonio, San Antonio, TX 78249; and [c]Department of Sociology and Criminology, The Pennsylvania State University, University Park, PA 16802

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved April 20, 2020 (received for review February 27, 2020)

The application of a currently proposed differential privacy algorithm to the 2020 United States Census data and additional data products may affect the usefulness of these data, the accuracy of estimates and rates derived from them, and critical knowledge about social phenomena such as health disparities. We

more is known about the impact of DP on important uses of census data, particularly in light of concerns discussed at the workshop on December 2019 (4).

Why are there concerns about privacy? Title 13 of the United States Code imposes heavy obligations on the US Census Bureau

Findings of recent publications include:

- Systematically undercounts the population in mixed-race and mixed-partisan precincts.
- Introduces substantial distortion in COVID-19 mortality rates, sometimes causing mortality rates to exceed 100 percent.
- Net migration estimates by five-year age groups would only be accurate enough for use in about half of counties.

Conclusions

- The Census Bureau is now in the midst of a radical escalation of disclosure control justified by an unsuccessful database reconstruction experiment.
- The actual threat posed by the Census Bureau's reconstruction is similar to the threat posed by randomly guessing respondent characteristics.

Conclusions

- The database reconstruction experiment failed to demonstrate a disclosure threat and cannot justify degradation of the nation’s statistical infrastructure.
- Reducing data quality is likely to backfire, as it will erode public trust and weaken the core justification for census data collection

Thank you.

Questions?