Innovating Data Privacy for the American Community Survey

Rolando A. Rodríguez and Amy D. Lauger
Center for Enterprise Dissemination – Disclosure Avoidance
U.S. Census Bureau
2019 ACS Data Users Conference

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

The views in this presentation are those of the authors and not those of the U.S. Census Bureau.

The Census Bureau is committed to data quality

- The Census Bureau's *mission* is to serve as the nation's leading provider of **quality** data about its people and economy.
- The Census Bureau's *goal* is to provide the best mix of timeliness, relevancy, **quality** and cost for the data we collect and services we provide.
- The American Community Survey (ACS) helps local officials, community leaders, and businesses understand the changes taking place in their communities. It is the **premier source** for detailed population and housing information about our nation.

The Census Bureau is also committed to data privacy protection

- The Census Bureau operates, collects data, and publishes statistics under the authority of several titles of the U.S. Code
- Title 13, Sec.9: Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof […] may […] make any publication whereby the **data furnished by any particular establishment or individual** under this title **can be identified**
- ACS response requirement is tied to guarantee of **confidentiality**
- Methods that protect **confidentiality/privacy** are vital for public data releases

Protect the person. Describe the people.

- ACS data products must be statistical rather than personal
- Public data releases contain information about each respondent
- Disclosure avoidance (DA) methods are used to control this risk
- Previous ACS products have used a number of DA methods

Internal Files	Public Tables	Public-Use Microdata
Household Swapping	Category Collapsing	Geographic aggregation
Synthetic Data		Sub-sampling
		Variable suppression
		Category collapsing
		Variable top-coding

We must innovate DA for the ACS while upholding quality

- Current DA methods are based on assumptions that no longer hold
- The Census Bureau's Data Stewardship Executive Policy committee has charged the ACS with improving its methods
- We are actively researching methods that will allow for provable privacy guarantees
- The ACS will use these methods once fully developed

First, we will consider expanding current DA methods

- The Census Bureau intends to release the standard complement of ACS tables and public-use microdata (PUMS) for 2018
- The detail in published tables and in the PUMS are major disclosure avoidance concerns and are co-dependent
- We will assess expanding the use of synthetic data to protect the PUMS
- Subject matter experts at the Census Bureau will evaluate the accuracy of new synthetic data methods

Then, ACS will transition to formal privacy methods

- Formal privacy methods allow policy makers to balance important social goods – data accuracy and data privacy
- Data users are already familiar with such tradeoffs and accounting for them
  - Complex sample design and survey weights to account for it
  - Effect of new DA methods will be another source of published error
- More transparency about DA methods and their effects will be possible
- We must have the involvement of data users to understand the effects of various options fully

Many experts are advising the Census Bureau as we modernize disclosure avoidance

- The Census Bureau has engaged academic and industry partners with expertise in modern data privacy methods
- Teams of experts are working with Census Bureau staff on multiple Census Bureau products, including 2020 Census, ACS and 2017 Economic Census
- The Census Bureau's subject matter experts are closely monitoring the effects of new methods on ACS accuracy
- The Census Bureau has actively sought feedback from the data user community

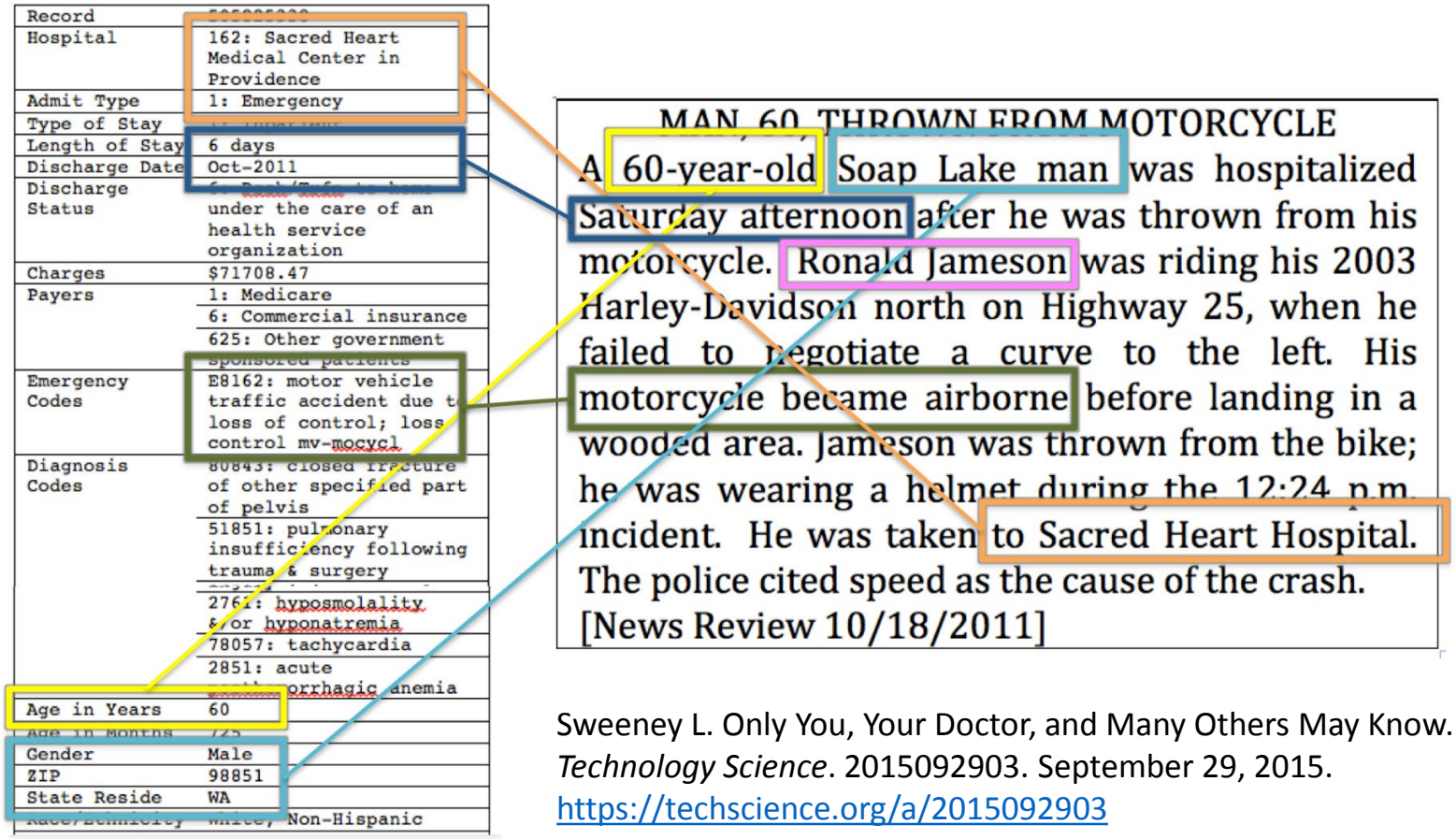
DA methods must adapt to a new data environment

- Data curators face an imposing global data environment
- Many sources of detailed personal information now readily available
- Modern computational methods designed to harness these sources
- ACS must maintain respondent privacy within this atmosphere
- Re-identification risks are an established reality, not merely a theoretical concept

In 1997, a governor had his medical records identified

- In 1996, the governor of Massachusetts collapsed at a keynote address
- After recovery, he oversaw the public release of hospital data for use in improving healthcare and related costs and assured its privacy
- Respondent privacy was protected by commonly used de-identification techniques
- For $20, an MIT graduate student purchased voter rolls, which included name, address, ZIP, birthdate, and sex
- She re-identified the governor's records in the hospital data and sent him a copy

In 2018, computational power makes similar attacks easier and more widespread



Record	
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government
Emergency Codes	E8162: motor vehicle traffic accident due t loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute orrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
	Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

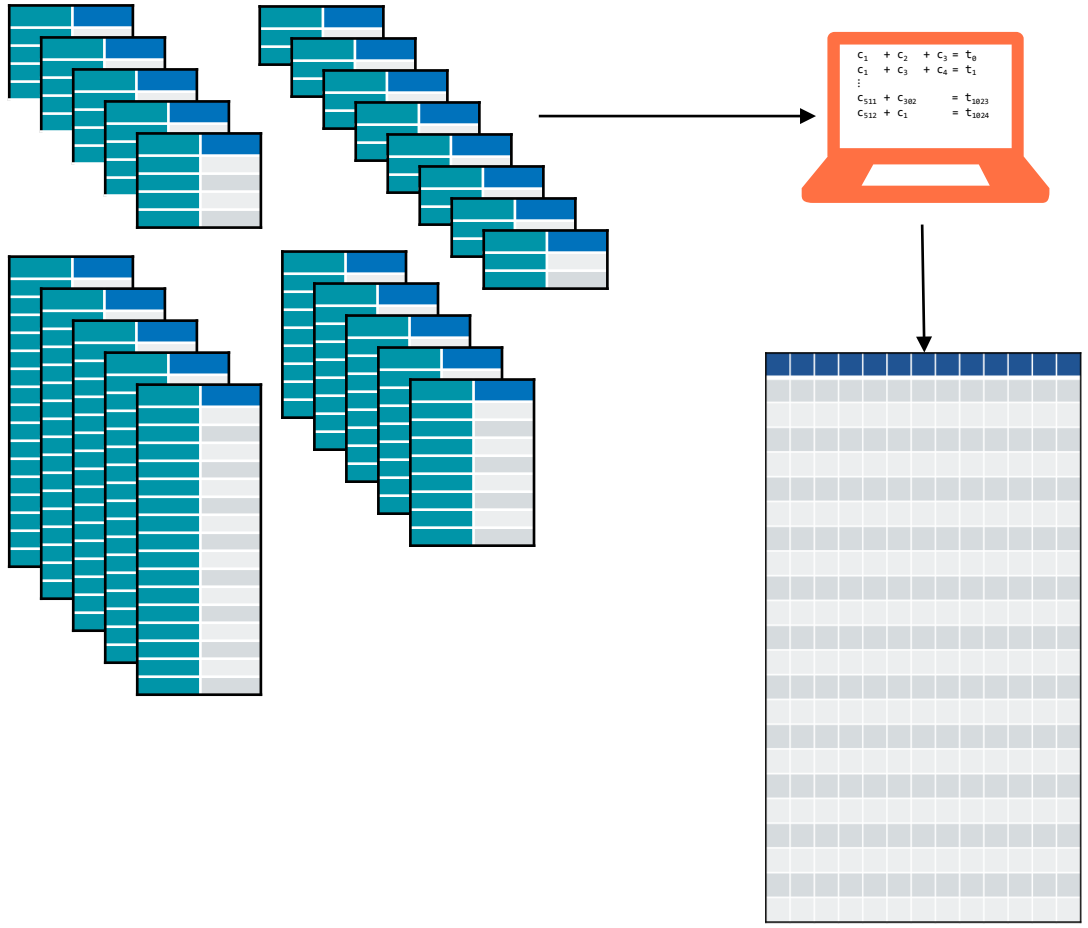
Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. 2015092903. September 29, 2015. https://techscience.org/a/2015092903

Several publicized re-identification attacks have happened

- Re-identification actually happens, with seemingly protected data
- Simple removal of obvious identifying information is not sufficient
- People can be unique on a relatively small set of attributes
- Current DA methods are not sufficient against known current risks
- Re-identification not limited to microdata releases
- Database reconstruction from public tables is possible

Database reconstruction turns statistics back into people

- Database reconstruction aims to recreate a non-public dataset from publicly available data
- Large table releases can easily yield multiple cells per item in the non-public file
- Modern software makes combining tables into microdata feasible
- Current ACS DA practices are not designed to protect from these types of attacks



The Census Bureau used 2010 Census tables to demonstrate database reconstruction

- Used public tables to create a database row for each person
- More than 50% of the population is unique on a combination of block, age, sex, race, and ethnicity (exact percentage is confidential)
- 45% of rows linkable to commercial data with personal information (138 million people)
- 38% of linked rows matched back to named Census records (52 million people)
- The overall vulnerability is 17% of population
- Previous re-identification study had overall vulnerability of 0.0038%

Public ACS data creates its own privacy risks

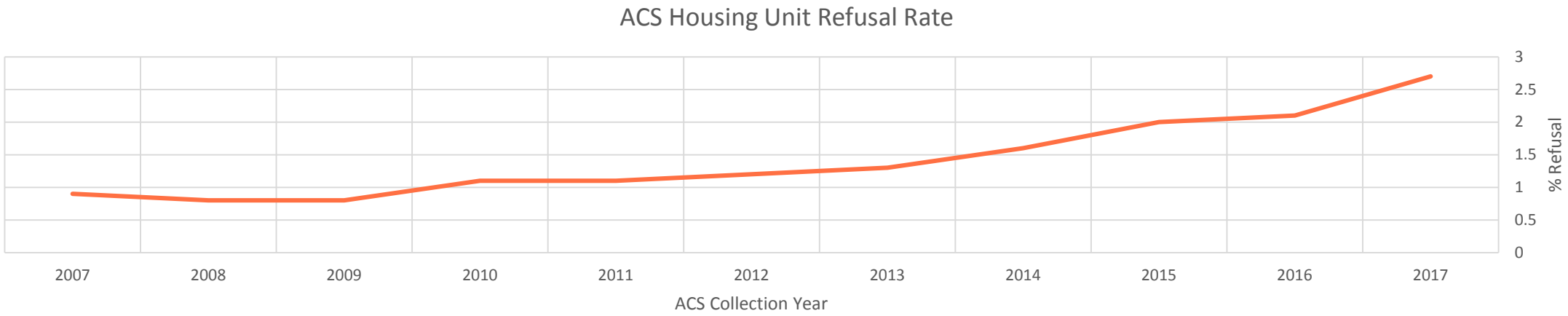
- The Census Bureau publishes multiple ACS data products
  - Summary tables
  - Custom tabulations
  - Public-use microdata samples (PUMS)
- The main DA method for ACS PUMS is limiting geography
  - The smaller the geography, the more likely someone is unique
  - Smallest geographic unit in PUMS is the public-use microdata area (PUMA)
  - PUMAs must contain populations of at least 100,000 people
- What if someone can break PUMAs into smaller geographies?

ACS tables are the key to unlocking PUMAs

- Information on smaller geographies already available from ACS tables
- ACS releases hundreds of 5-year tables by block group
- Innovative statistical models have combined PUMS and tables to make new sub-PUMA estimates
- Although not the intent of the models, new estimates can add to reconstruction and re-identification risk
- ACS needs DA methods that can protect against this kind of risk… and other risks that may come

The Census Bureau needs to maintain trust of our respondents

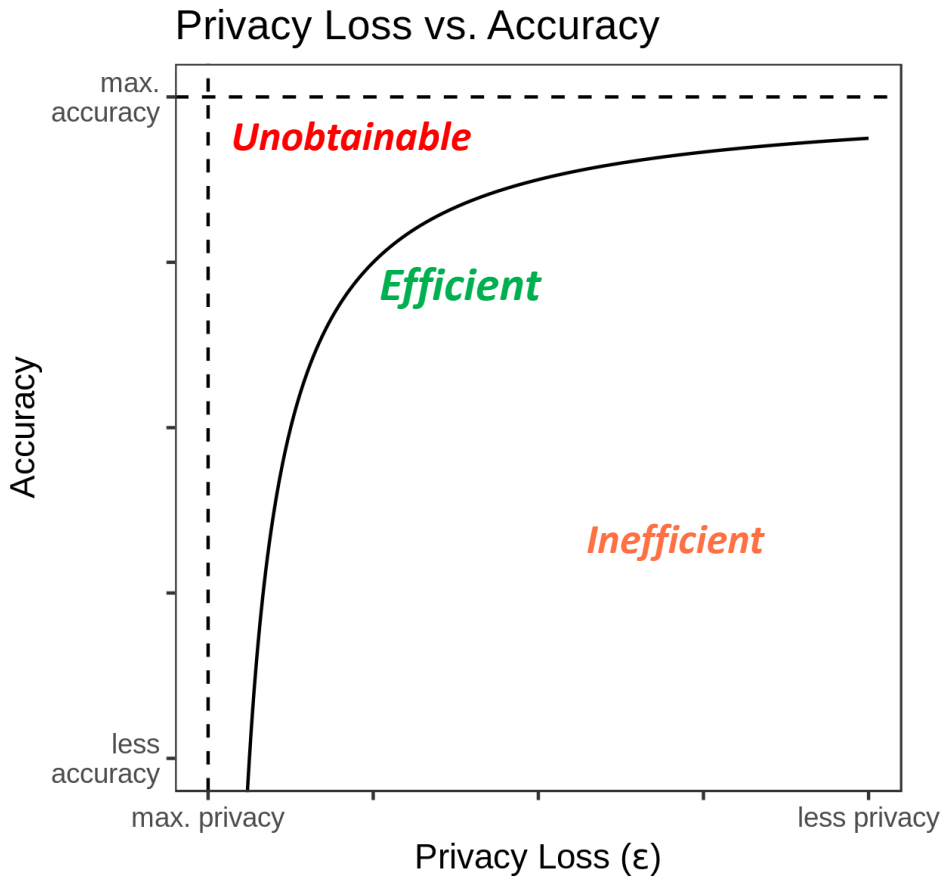
- People have growing concerns about their privacy
- ACS must maintain respondent trust within this environment
- Decreases in respondent trust could have negative impacts on quality
- Provable privacy guarantees will aid in keeping that trust



www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/

Formal privacy methods can future-proof the privacy of the ACS data

- Can guarantee privacy against broad classes of attacks
- Do not depend on which datasets are now or will be available
- Have a calculable, global privacy loss for a given set of releases at a given accuracy
- Allow for transparency about the method, data accuracy, and privacy loss



Formal privacy for the ACS is a work in progress

- Formal privacy methods developed for 2020 Census will inform ACS methods
- ACS has additional features that require further research
  - More topic areas
  - Complex survey design and survey weights
  - Population controls
- ACS will implement formal privacy methods once developed to have a satisfactory balance of data privacy and data usefulness
- Until then, ACS will expand non-formal methods to provide protection

New DA methods can also improve quality

- Some current DA methods may be reduced or removed
  - PUMS sub-sampling
  - PUMS geographic aggregation
  - Household swapping
  - Variable top-coding
  - Category collapsing
- New methods can work alongside other ACS innovations
  - Data modeling, especially for small areas and missing data
  - Use of administrative records
- Transparency of DA methods might allow data users to account for the effect of privacy protection in their analyses

Going forward together…

- The Census Bureau understands the importance of the ACS to the nation
- We understand many users will find these DA changes concerning
- The Census Bureau must act to fulfill our obligations to data users and to respondents
- Data user involvement is and will be vital
  - Provide input to the Census Bureau for data needs
  - Information sharing among peers for adapting to new public data
- These changes will let the Census Bureau better serve data users
  - Transparency
  - Data-driven decisions about balancing data privacy with data accuracy
  - Continue to have a quality, trusted, reputable product for years to come