Deploying Differential Privacy for the 2020 Census of Population and Housing

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau

July 16, 2019
Privacy Enhancing Technologies Symposium
Stockholm, Sweden 2019

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

Abstract

en differential privacy was created more than a decade ago, the motivating example was statistics published by an offici
tistics agency. In theory there is no difference between theory and practice, but in practice there is.

ttempting to transition differential privacy from the theory to practice, and in particular for the 2020 Census of Population
l Housing, the U.S. Census Bureau has encountered many challenges unanticipated by differential privacy's creators.
ny of these challenges had less to do with the mathematics of differential privacy and more to do with operational
uirements that differential privacy's creators had not discussed in their writings. These challenges included obtaining
alified personnel and a suitable computing environment, the difficulty of accounting for all uses of the confidential data, th
k of release mechanisms that align with the needs of data users, the expectation on the part of data users that they will
re access to micro-data, the difficulty in setting the value of the privacy-loss parameter, ε (epsilon), and the lack of tools
l trained individuals to verify the correctness of differential privacy, and push-back from same members of the data user
nmunity.

dressing these concerns required developing a novel hierarchical algorithm that makes extensive use of a high-
formance commercial optimizer; transitioning the computing environment to the cloud; educating insiders about different
acy; engaging with academics, data users, and the general public; and redesigning both data flows inside the Census
eau and some of the final data publications to be in line with the demands of formal privacy.

Acknowledgments

his presentation incorporates work by:

- Dan Kifer (Scientific Lead)
- John Abowd (Chief Scientist)
- Tammy Adams, Robert Ashmead, Aref Dajani, Jason Devine, Natha Goldschlag, Michael Hay, Cynthia Hollingsworth, Meriton Ibrahimi, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Ned Porter, Anne Ross, William Sexton, Lars Vilhuber, and Pavel Zhuravlev

Outline

- Motivation
- The flow of census response data
- Disclosure Avoidance for the 2010 census
- Disclosure Avoidance for the 2018 census End-to-End test
- Disclosure Avoidance for the 2020 census
- Conclusion

Motivation

The 2020 Census of Population and Housing

Count everyone once, only once, and in the right place.

Motivation
Article 1, Section 2

"…The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and withi every subsequent Term ten Years, in such Mann as they shall by Law direct…"

Article 1, Section 2

he House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in ach State shall have the Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

o Person shall be a Representative who shall not have attained to the Age of twenty five Years, and been seven Years a Citizen of the United tates, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

epresentatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their espective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of ears, and excluding Indians not taxed, three fifths of all other Persons. **The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.**

he Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at Least one Representative; and until uch enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode-Island and Providence lantations one, Connecticut five, New-York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, outh Carolina five, and Georgia three.

/hen vacancies happen in the Representation from any State, the Executive Authority thereof shall issue Writs of Election to fill such Vacancies.

he House of Representatives shall chuse their Speaker and other Officers; and shall have the sole Power of Impeachment.

"in such Manner as they shall by Law direct."
Public Law 94-171

Public Law 94-171
94th Congress

An Act

Dec. 23, 1975
[H.R. 1753]

To amend section 141 of title 13, United States Code, to provide for the transmittal to each of the several States of the tabulation of population of that State obtained in each decennial census and desired for the apportionment or districting of the legislative body or bodies of that State, in accordance with, and subject to the approval of the Secretary of Commerce, a plan and form suggested by that officer or public body having responsibility for legislative apportionment or districting of the State being tabulated, and for other purposes.

Population, tabulation for State legislative apportionment.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, That section 141 of title 13, United States Code, is amended by adding at the end thereof the following new subsection:

"(c) The officers or public bodies having initial responsibility for the legislative apportionment or districting of each State may, not later than three years prior to the census date, submit to the Secretary a plan identifying the geographic areas for which specific tabulations of population are desired. Each such plan shall be developed in accordance with criteria established by the Secretary, which he shall furnish to such officers or public bodies not later than April 1 of the fourth year preceding the census date. Such criteria shall include requirements which assure that such plan shall be developed in a nonpartisan manner. Should the Secretary find that a plan submitted by such officers or public bodies does not meet the criteria established by him, he shall consult to the extent necessary with such officers or public bodies in order to achieve the alterations in such plan that he deems necessary to bring it into accord with such criteria. Any issues with respect to such plan remaining unresolved after such consultation shall be resolved by the Secretary, and in all cases he shall have final authority for determining the geographic format of such plan. Tabulations of population for the areas identified in any plan approved by the Secretary shall be completed by him as expeditiously as possible after the census date and reported to the Governor of the State involved and the officers or public bodies having responsibility for legislative apportionment or districting of such State, except that such tabulations of population of each State requesting a tabulation plan, and basic tabulations of population of each other State, shall, in any event, be completed, reported and transmitted to each respective State within one year after the census date.".

SEC. 2. (a) The heading for section 141 of title 13, United States Code, is amended by adding at the end thereof the following: "**; tabulation for legislative apportionment**".

(b) The table of sections for chapter 5 of title 13, United States Code, is amended by striking out the item relating to section 141 and inserting in lieu thereof the following:

"141. Population, unemployment, and housing; tabulation for legislative apportionment.".

Approved December 23, 1975.

LEGISLATIVE HISTORY:

HOUSE REPORT No. 94-456 (Comm. on Post Office and Civil Service).
SENATE REPORT No. 94-539 (Comm. on Post Office and Civil Service).
CONGRESSIONAL RECORD, Vol. 121 (1975):
Nov. 7, considered and passed House.
Dec. 15, considered and passed Senate.

http://uscode.house.gov/statutes/pl/94/171.pdf

Federal Register / Vol. 82, No. 215 / Nov 8, 2017 / Notices (for the 2018 End-to-End test)

Dec. 31, 2018

We will report (per block):

P1. RACE/ETHNICITY
Universe: Total population
Group by: BLOCK

P2. RACE/ETHNICITY
Universe: Total population age 18 and over

H1. OCCUPANCY STATUS

P42. GROUP QUARTERS POPULATION
Universe: Population in Group Quarters

DEPARTMENT OF COMMERCE

Bureau of the Census

[Docket Number 170824806–7806–01]

Proposed Content for the Prototype 2020 Census Redistricting Data File

AGENCY: Bureau of the Census, Department of Commerce.

ACTION: Notice and request for comment.

SUMMARY: The 2020 Census Redistricting Data Program provides states the opportunity to specify the small geographic areas for which they wish to receive 2020 decennial population totals for the purpose of reapportionment and redistricting. This notice pertains to Phase 3, the Data Delivery phase of the program, as the U.S. Census Bureau is providing notification and requesting comment on the content of the prototype 2020 Census Redistricting Data File that will be produced from the 2018 End-to-End Census Test. The Census Bureau anticipates publishing the content for the prototype 2020 Census Redistricting Data File from the 2018 End-to-End Census Test in the second quarter of fiscal year 2018 in a final notice. In that final notice, the Census Bureau also will respond to the comments received on this notice.

We need to protect privacy!
13 U.S. Code § 9 - Information as confidential; exception

(a) **Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison may**, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1998.

(1) Use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(2) Make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3) Permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports. No department, bureau, agency, officer, or employee of the Government, except the Secretary in carrying out the purposes of this title, shall require, for any reason, copies of census reports which have been retained by any such establishment or individual. **Copies of census reports, which have been so retained, shall be immune from legal process, and shall not, without the consent of the individual or establishment concerned, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding.**

(b) The provisions of subsection (a) of this section relating to the confidential treatment of data for particular individuals and establishments, shall not apply to the censuses of governments provided for by subchapter III of chapter 5 of this title, nor to interim current data provided for by subchapter IV of chapter 5 of this title as to the subjects covered by censuses of governments, with respect to any information obtained therefore that is compiled from, or customarily provided in, public records.

Statistical agencies collect data under a *pledge of confidentiality.*

We pledge:

- Collected data will be used *only for statistical purposes*.
- Collected data will be kept *confidential*.
- Data from individuals or establishments *won't be identifiable in any publication.*

Disclosure Avoidance for the 2010 Census

IT'S IN OUR HANDS

United States Census 2010

“Disclosure Avoidance”
means
preventing improper *disclosures*.

This is the
fficial form for
ll the people at
his address.”

United States Census 2010

This is the official form for all the people at this address.
It is quick and easy, and your answers are protected by law.

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

Use a blue or black pen.

Start here

The Census must count every person living in the United States on April 1, 2010.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2010.
- Leave these people off your form, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

The Census must also include people without a permanent place to stay, so:

- If someone who has no permanent place to stay is staying here on April 1, 2010, count that person. Otherwise, he or she may be missed in the census.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?

Number of people = ☐

2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1? *Mark* ☒ *all that apply.*

☐ Children, such as newborn babies or foster children
☐ Relatives, such as adult children, cousins, or in-laws
☐ Nonrelatives, such as roommates or live-in baby sitters
☐ People staying here temporarily
☐ No additional people

3. Is this house, apartment, or mobile home — *Mark* ☒ *ONE box.*

☐ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
☐ Rented?
☐ Occupied without payment of rent?

4. What is your telephone number? *We may call if we don't understand an answer.*

Area Code + Number

OMB No. 0607-0919-C: Approval Expires 12/31/2011.

Form **D-61** (9-25-2008)

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.
What is Person 1's name? *Print name below.*

Last Name

First Name MI

6. What is Person 1's sex? *Mark* ☒ *ONE box.*

☐ Male ☐ Female

7. What is Person 1's age and what is Person 1's date of birth? *Please report babies as age 0 when the child is less than 1 year old. Print numbers in boxes.*

Age on April 1, 2010 Month Day Year of birth

→ NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

☐ No, not of Hispanic, Latino, or Spanish origin
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark* ☒ *one or more boxes.*

☐ White
☐ Black, African Am., or Negro
☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

☐ Asian Indian ☐ Japanese ☐ Native Hawaiian
☐ Chinese ☐ Korean ☐ Guamanian or Chamorro
☐ Filipino ☐ Vietnamese ☐ Samoan
☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

☐ Some other race — *Print race.*

10. Does Person 1 sometimes live or stay somewhere else?

☐ No ☐ Yes — *Mark* ☒ *all that apply.*

☐ In college housing ☐ For child custody
☐ In the military ☐ In jail or prison
☐ At a seasonal or second residence ☐ In a nursing home
☐ For another reason

→ If more people were counted in Question 1, continue with Person 2.

“It is quick an
easy, and you
answers are
protected by
law.”

The 2010 Census collected data on 308,745,538 people.

DRF

aw data from
espondents:
Decennial
esponse File

Statistical Processes

HDF

Hundred-percent Detail File

Pre-specifiec
tabular
summaries:
PL94-171, SF
SF2 (SF3, SF
… in 2000)

Special tabulations a
post-census
research

Confidential data

Public data

The 2000 and 2010 Confidentiality Mechanism operated as a filter on the Census Edited File:

DRF

w data from
spondents:
ecennial
sponse File

Selection & unduplication:
Census Unedited File

Edits, imputations:
Census Edited File

Confidentiality Mechanism

HDF

Hundred-percent Detail File

Pre-specified tabu
summaries:
PL94-171, SF1, S
(SF3, SF4, … in 2

Special tabulatio
and post-censu
research

Public data

For each person, we collected 6 variables (44 bits of data)

Variable	Range	Bits
Block	6,207,027 inhabited blocks	23
Sex	2 (Female/Male)	1
Age	103 (0-99 single age year categories, 100-104, 105-109, 110+)	7
Race	63 allowable race combinations	7
Ethnicity	2 (Hispanic/Not)	1
Relationship	17 values	5
Total		44

308,745,538 people x 6 variables = 1,852,473,228 measuremen

308,745,538 people x 44 bits = 13,584,803,672 bits ≈ 1.7 GB

2010 Census: Summary of Publications (approximate counts)

Publication	Released counts
PL94-171 Redistricting	2,771,998,263
Balance of Summary File 1	2,806,899,669
Summary File 2	2,093,683,376
Public-use micro data sample	30,874,554
Lower bound on published statistics	7,703,455,862
Published Statistics/person	25
Recall: Collected variables/person:	6
Published Statistics/collected variable	$25 \div 6 \approx 4.2$

Question:
s it possible to run the statistical process in reverse?

DRF

aw data from
espondents:
Decennial
esponse File

Reconstruction Processes

Pre-specified
tabular
summaries:
PL94-171,
SF1, SF2

Special
tabulations ar
post-census
research

Public data

1,852,473,228 collected values

(308,745,538 people x 6 variables/person)

2003: Database Reconstruction

ABSTRACT

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an n-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to be answered by $\sum_{i \in q} d_i$. Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude $\Omega(\sqrt{n})$. That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude $\tilde{O}(\sqrt{n})$.

For time-T bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{T}$.

Revealing Information while Preserving Privacy

Irit Dinur Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

ABSTRACT

We examine the tradeoff between privacy and usability of

research which is based (among other things) on statistics of the information in the database. On the other hand, the hospital is obliged to keep the privacy of its patients, i.e. leak no medical information that could be related to a specific patient. The hospital needs an access mechanism to the database that allows certain 'statistical' queries to be answered, as long as they do not violate the privacy of any single patient.

*Work partly done when the author was at DIMACS, Rutgers University, and while visiting Microsoft Research Silicon Valley Lab.

PODS 2003, June 9-12, 2003, San Diego, CA.

One simple tempting solution is to remove from the database all 'identifying' attributes such as the patients' names and social security numbers. However, this solution is not enough

In their comparative survey of privacy methods for statistical databases, Adam and Wortmann [2] classified the approaches taken into three main categories: (i) query restriction, (ii) data perturbation, and (iii) output perturbation. We give a brief review of these approaches below, and refer the reader to [2] for a detailed survey of the methods and their weaknesses.

Query Restriction. In the query restriction approach, queries are required to obey a special structure, supposedly to prevent the querying adversary from gaining too much information about specific database entries. The limit of this approach is that it allows for a relatively small number of queries.

A related idea is of query auditing [7], i.e. a log of the queries is kept, and every new query is checked for possible compromise, allowing/disallowing the query accordingly.

[1] A patient's gender, approximate age, approximate weight, ethnicity, and marital status – may already suffice for a complete identification of most patients in a database of a thousand patients. The situation is much worse if a relatively 'rare' attribute of some patient is known. For example, a patient having Cystic Fibrosis (frequency $\approx 1/3000$) may be uniquely identified within about a million patients.

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

Attacking statistical databases

Statistical agencies are trusted curators.

espondents

Confidential Database

Age Sex Race/MS
8 FBS
18 MWS
24 FWS
30 MWM
36 FBM
66 FBM
84 MBM

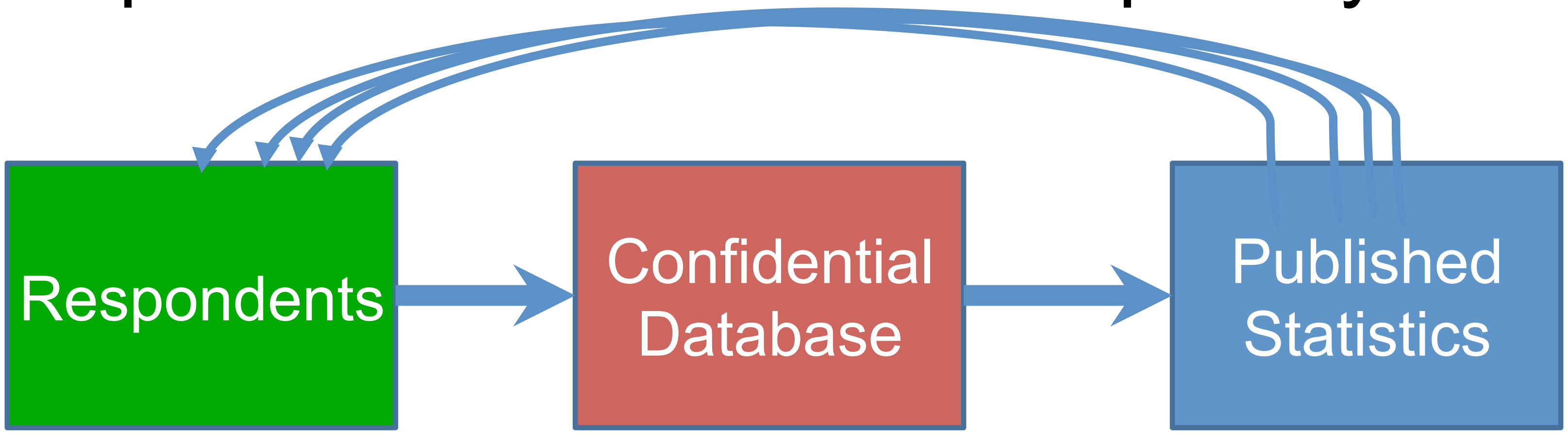
Published Statistic

	#	Median Age	Me A
Total	7	30	3
Women	4	30	33
Male	3	30	4
Black	4	51	48
White	3	24	2
Married	4	51	5
Black F	3	36	36

This is the trusted curator model

We now know “trusted curator” model is more complex.

Every data publication results in some privacy loss.



Publishing too many statistics results in the compromise of the entire confidential database.

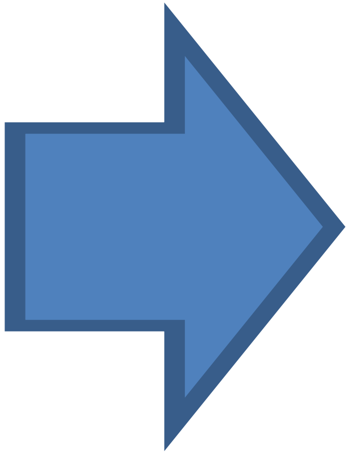
Consider the statistics from a single household

s Female White Single (24 FWS)

	Count	Median	Mean
Total	1	24	24
# Female	1	24	24
# white	1	24	24
Single	1	24	24
White F	1	24	24

Publishing statistics for this household alone would result in an improper disclosure.

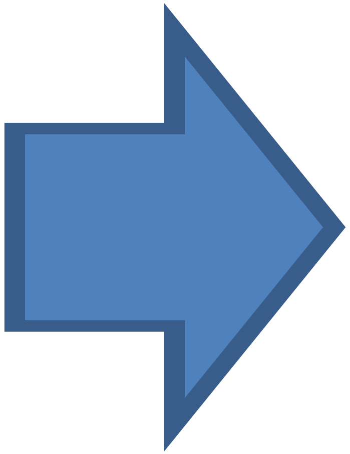
s Female White Single (24 FWS)



	Count	Median	Mean
Total	(D)	(D)	(D)
# Female	(D)	(D)	(D)
# white	(D)	(D)	(D)
Single	(D)	(D)	(D)
White F	(D)	(D)	(D)

(D) Means suppressed to prevent an improper disclosure

In the past, statistical agencies aggregated data from many households together into a single publication.



	Count	Median Age	Mean Age
Total	7	30	38
# Female	4	30	33.5
# male	3	30	44
# black	4	51	48.5
# white	3	24	24
Married	4	51	54
Black F	3	36	36.7

We now know that this publication can be reverse-engineered to reveal the confidential database.

66 FBM & 84 MBM

30 MWM & 36 FBM

8 FBS

18 MWS

24 FWS

	Count	Median	Mean
Total	7	30	38
# Female	4	30	33.5
# male	3	30	44
# black	4	51	48.5
# white	3	24	24
Married	4	51	54
Black F	3	36	36.7

This table can be expressed by 164 equation
Solving those equations takes
0.2 seconds on a 2013 MacBook Pro.

The problem with publishing fewer statistics:
t’s hard to know how many statistics is “too many.”

Solution #1	Solution #2
8 FBS	2 FBS
18 MWS	12 MWS
24 FWS	24 FWS
30 MWM	30 MBM
36 FBM	36 FWM
66 FBM	72 FBM
84 MBM	90 MBM

	Count	Median	Mean
Total	7	30	38
# Female	4	30	33.5
# male	3	30	44
# black	4	51	48.5
# white	3	24	24
Married	4	51	54
Black F	3	36	36.7

Here's what the system looks like:

Variables:

```
#define FEMALE 0
#define MALE   1
(int S1 FEMALE MALE)
(int S2 FEMALE MALE)
(int S3 FEMALE MALE)
(int S4 FEMALE MALE)
(int S5 FEMALE MALE)
(int S6 FEMALE MALE)
(int S7 FEMALE MALE)
```

Constraints:

```
;; there are three males
(= (+ (if (= S1 MALE) 1 0)
      (if (= S2 MALE) 1 0)
      (if (= S3 MALE) 1 0)
      (if (= S4 MALE) 1 0)
      (if (= S5 MALE) 1 0)
      (if (= S6 MALE) 1 0)
      (if (= S7 MALE) 1 0)
      )
   3)
```

2010 Census: Summary of Publications (approximate counts)

Publication	Released counts
PL94-171 Redistricting	2,771,998,263
Balance of Summary File 1	2,806,899,669
Summary File 2	2,093,683,376
Public-use micro data sample	30,874,554
Lower bound on published statistics	7,703,455,862
Published Statistics/person	25
Recall: Collected variables/person:	6
Published Statistics/collected variable	$25 \div 6 \approx 4.2$

We performed a database reconstruction and re-identification attack for all 308,745,538 people in 2010 Census

1. Reconstructed 308,745,538 microdata records.
2. Used 4 commercial databases of the 2010 US population acquired 2009-2011 in support of the 2010 Census

   Commercial database had: **NAME, ADDRESS, AGE, SEX**
3. Linked reconstructed records to the commercial database records

   Linked database has: **NAME, ADDRESS, AGE, SEX , ETHNICITY & RAC**

   Link rate: 45%
4. Compared linked database to Census Bureau confidential data

   Question: How often did the attack get the all variables including race and ethnicity right?

   Answer: 38% (17% of US population)

Our attack is good, but not perfect. An outside attacker would have a harder time.

We confirmed re-identification of 38% (17% of US population)

We did not reconstruct families.

We did not recover detailed self-identified race codes

An outside attacker:

Would not know which re-identifications are correct.

An outside attacker would need to confirm with another external data source.

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

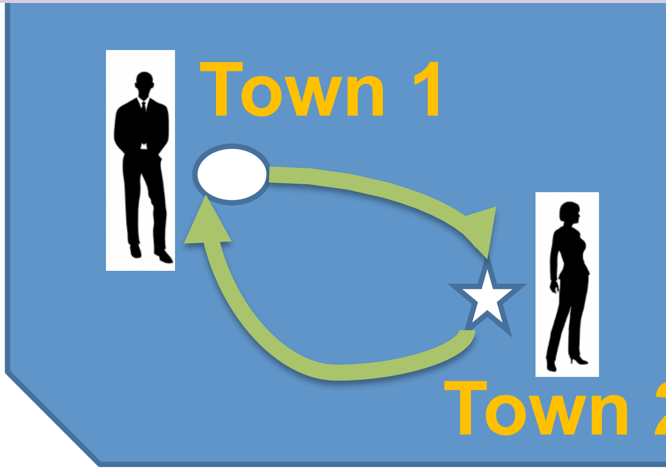
Some households were swapped with other households

- 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.

Confidentiality Mechan



Any State

Swapping is called a "Disclosure Avoidance" technique. It's job is to prevent improper disclosures.

We now know that the disclosure avoidance techniques we used in the 2010 Census were flawed

We released *exact population counts* for blocks, tracts and counties.

We did not swap 100% of the households

We released ≈25 statistics per person,
but only collected six pieces of data per person:

Block • Age • Sex • Race • Ethnicity • Relationship to Householder

The Census Bureau did the best possible in 2010.
The math for protecting a decennial census using formal privacy did not [yet] exist.

Faced with "database reconstruction," statistical agencies have just two choices

Option #1: Publish fewer statistics.

Option #2: Publish statistics with less accuracy.

Faced with “database reconstruction,” statistical agencies have just ~~two~~ one choice

~~Option #1: Publish fewer statistics.~~

Option #2: Publish statistics with less accuracy.



2006: Differential Privacy

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0, 1]$. We extend the study to general functions f, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function f. Roughly speaking, this is the amount that any single argument to f can change its output. The new analysis shows that for several particular applications substantially less noise is needed than was previously understood to be the case.

The first step is a very clean characterization of privacy in terms of indistinguishability of transcripts. Additionally, we obtain separation results showing the increased value of interactive sanitization mechanisms over non-interactive.

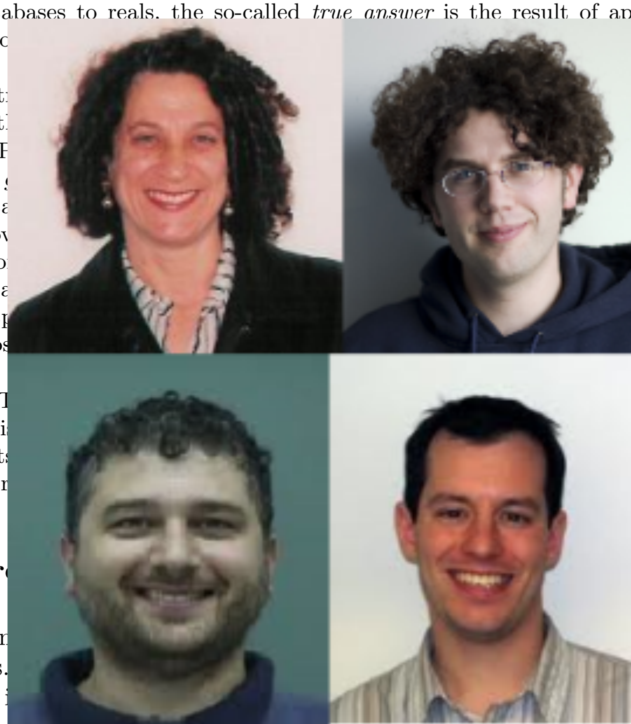
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3*]

[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying



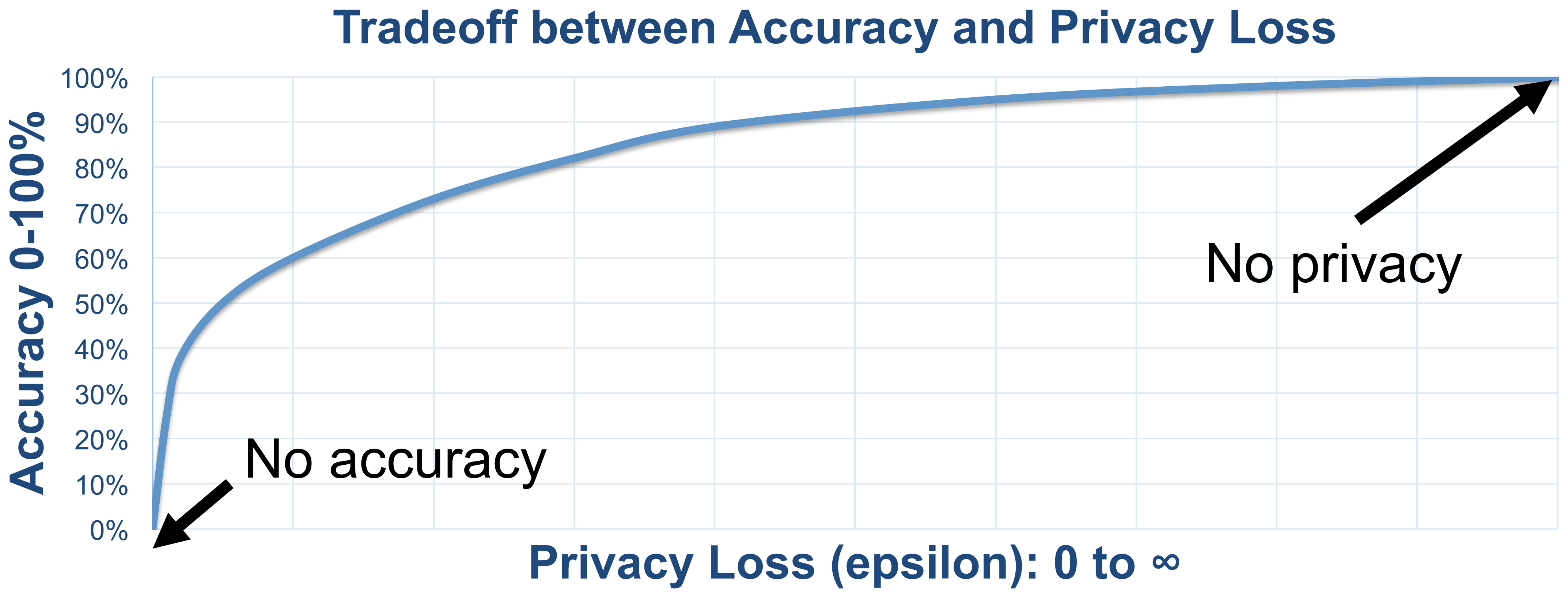
1 Intr

a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole while protecting the privacy of the individual contributors.

We assume the database is held by a trusted server. On input a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition

* Supported by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

Differential privacy gives us a mathematical approach for balancing accuracy and privacy loss



“Differential privacy” is really two things

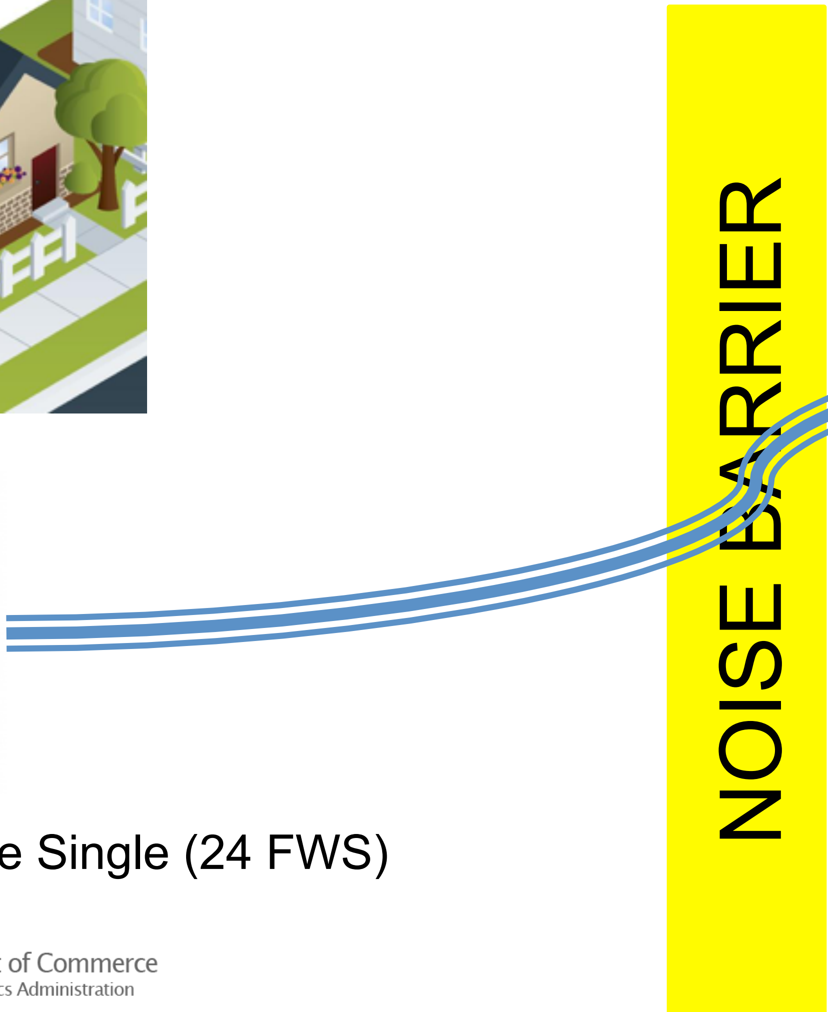
1 – A mathematical definition of privacy loss.

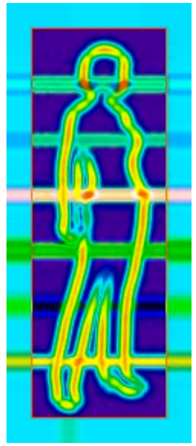
2 – Specific mechanisms that allow us to:

- ✓ *Add the smallest amount of noise necessary for a given privacy outcome*
- ✓ *Structure the noise to have minimal impact on the more important statistics*

Differential privacy — the big idea:
Use “noise” to create uncertainty about private data

s Female White Single (24 FWS)





35 yrs Female Black Single (35 F

Impact of the noise ≈ impact of a single pers

Impact of noise on aggregate statistics
decreases with larger population.

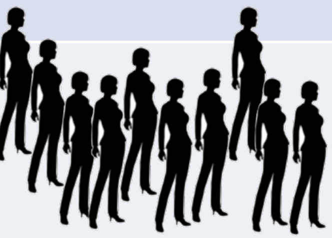
Each time we go through the noise barrier, we get a different number

		Age	Count
 person age 22	NOISE BARRIER	25	3
 1 person age 22		17	1
 1 person age 22		27	-1

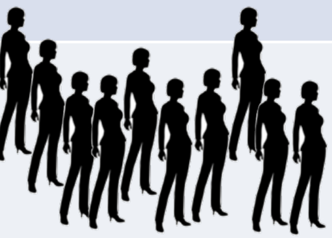
Epsilon controls the amount of noise

		Epsilon	Age
 person age 22	NOISE BARRIER	100	22
 person age 22		1.0	24
 person age 22		0.1	-115

Understanding the impact of "noise:" Statistics based on 10,000 experiments, epsilon=1.0

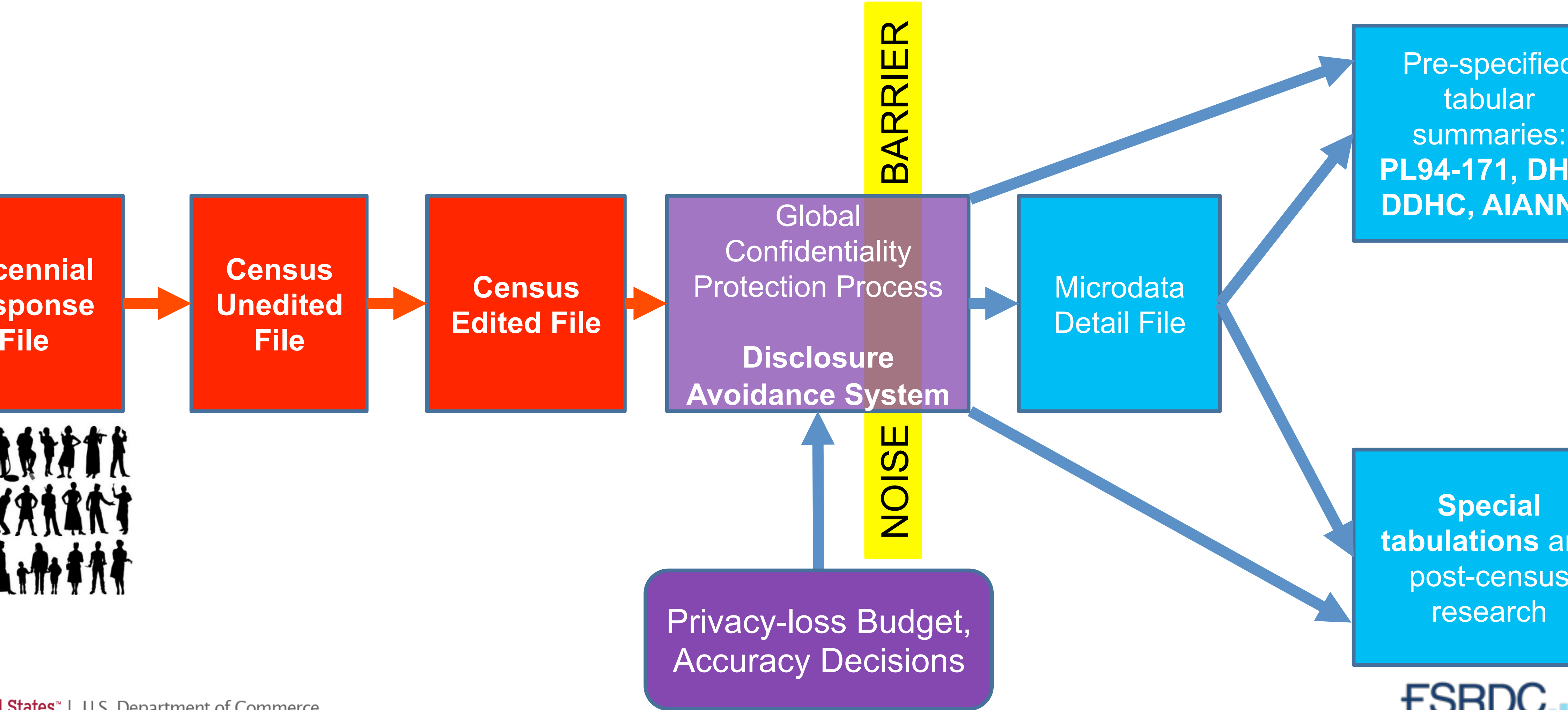
	NOISE BARRIER	5,000 (50%) runs	9,500 (95%) runs
 person age 22		Median(age): 9 → 73	Median(age): 0→ 104
 0 people, all age 22		Median(age): 17 → 61	Median(age): 0→ 103
 00 people, all age 22		Median(age): 21 → 22	Median(age): 21→ 22

The noise also impacts the person counts

		5,000 (50%) runs	9,500 (95%) runs
 person age 22	NOISE BARRIER	Median(age): 9 → 73 # people: -9 → 11	Median(age): 0→ 104 # people: -29 → 30
 0 people, all age 22		Median(age): 17 → 61 # people: 0 → 20	Median(age): 0→ 103 # people: -19 → 38
 00 people, all age 22		Median(age): 21 → 22 # people: 90 → 110	Median(age): 21→ 22 # people: 71 → 129

The 2020 census and differential privacy

DAS allows the Census Bureau to enforce global confidentiality protections

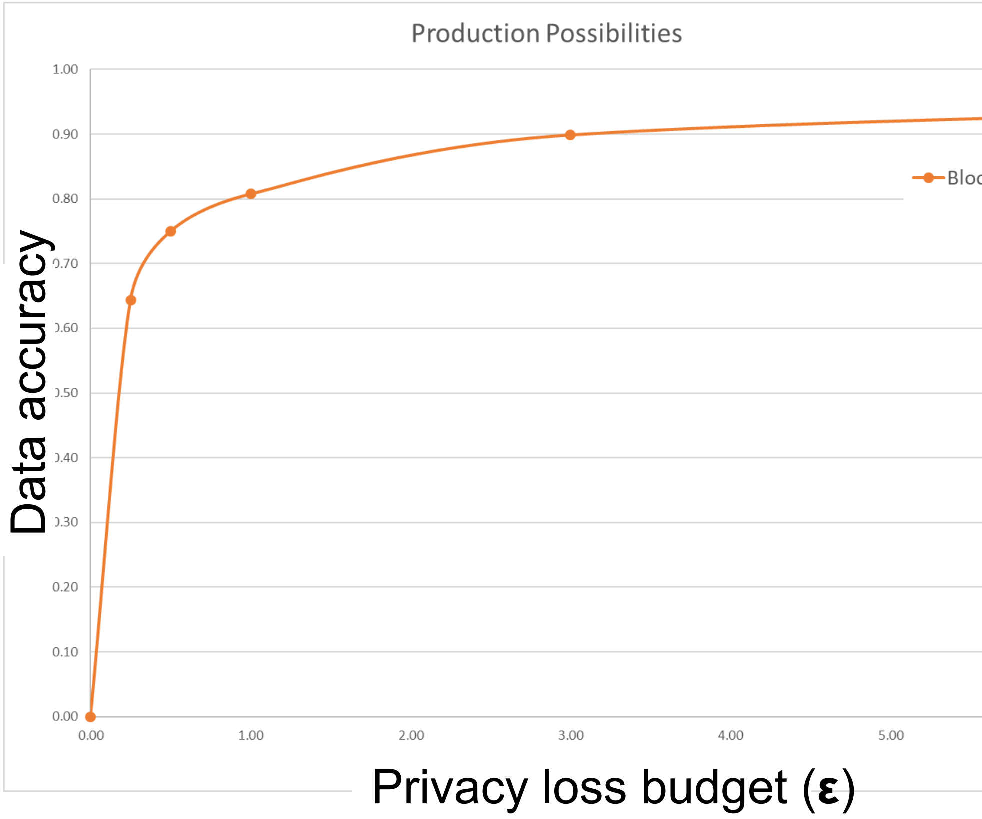


The Census disclosure avoidance system uses differential privacy to defend against an accurate reconstruction attack

Differential privacy provides:

Provable bounds on the accuracy of the best possible database reconstruction given the released tabulations.

Algorithms that allow policy makers to decide the trade-off between accuracy and privacy.

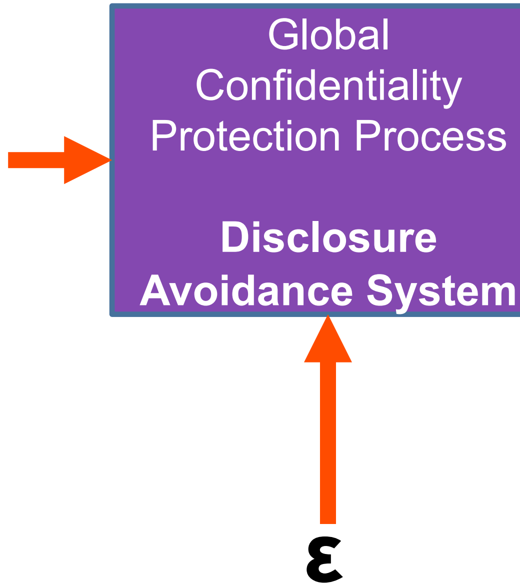


Final privacy-loss budget determined by Data Stewardship Executive Policy Committee (DSEP) with recommendation from Disclosure Review Board (DRB)

The Disclosure Avoidance System relies on injects formally private noise

Advantages of noise injection with formal privacy:

- Transparency: the details can be explained to the public
- Tunable privacy guarantees
- Privacy guarantees do not depend on external data
- Protects against accurate database reconstruction
- Protects every member of the population



Challenges:

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

There was no off-the-shelf system for applying differential privacy to a national census

We had to create a new system that:

- Produced higher-quality statistics at more densely populated geographies
- Produced consistent tables

We created new differential privacy algorithms and processing systems that:

- Produce highly accurate statistics for large populations (e.g. states, counties)
- Create protected microdata that can be used for any tabulation without additional privacy loss
- Fit into the decennial census production system

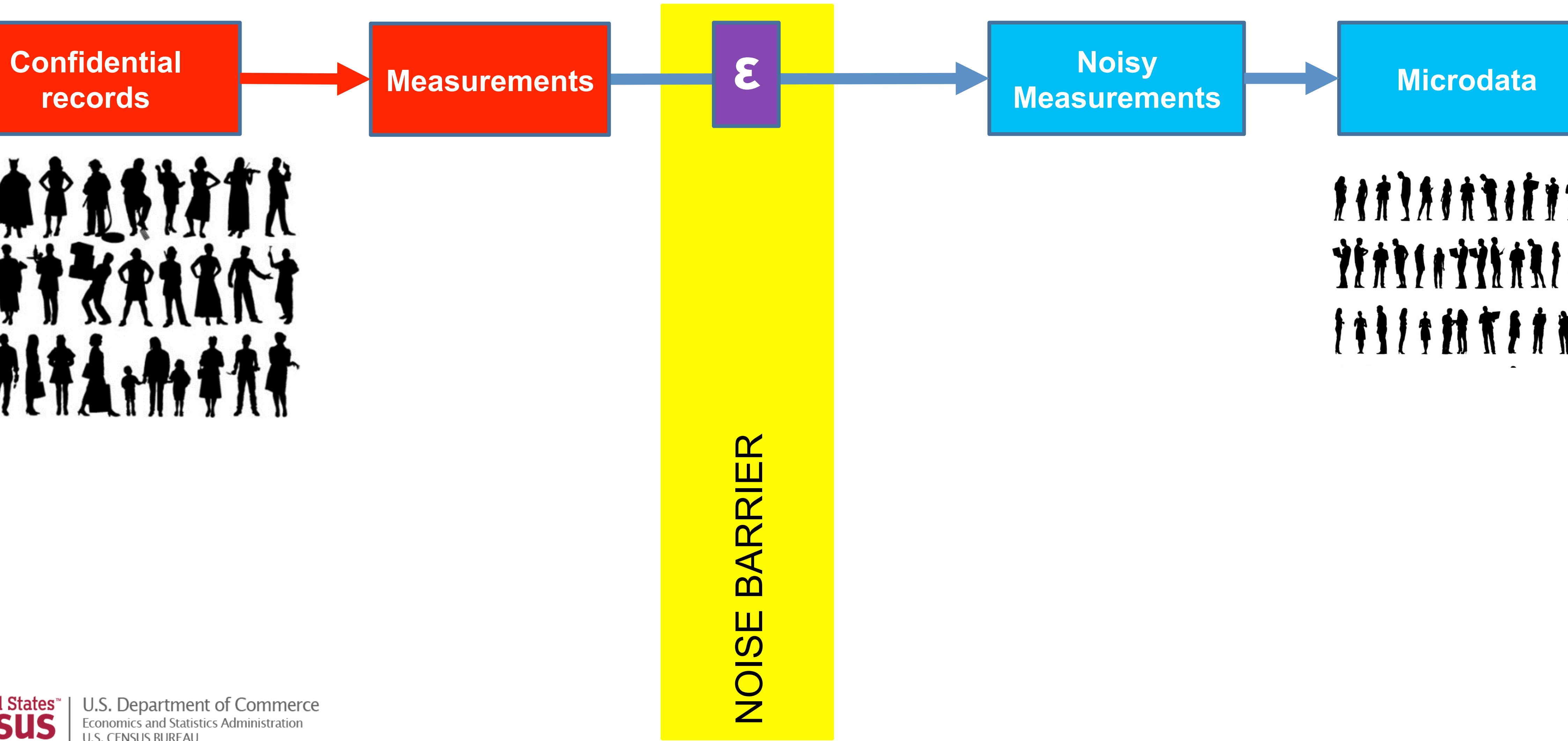
How the 2020 System Works: High-level Overview

- Every record in the population may be modified
  - *But modifications are bounded by the global privacy budget.*
- Records in the tabulation data have no exact counterpart in the confidential data
  - *There is no one-to-one mapping between CEF and MDF records.*
- Explicitly protected tabulations (PL-94 and SF-1) have provable, public accuracy levels

Basic approach for a DP Census

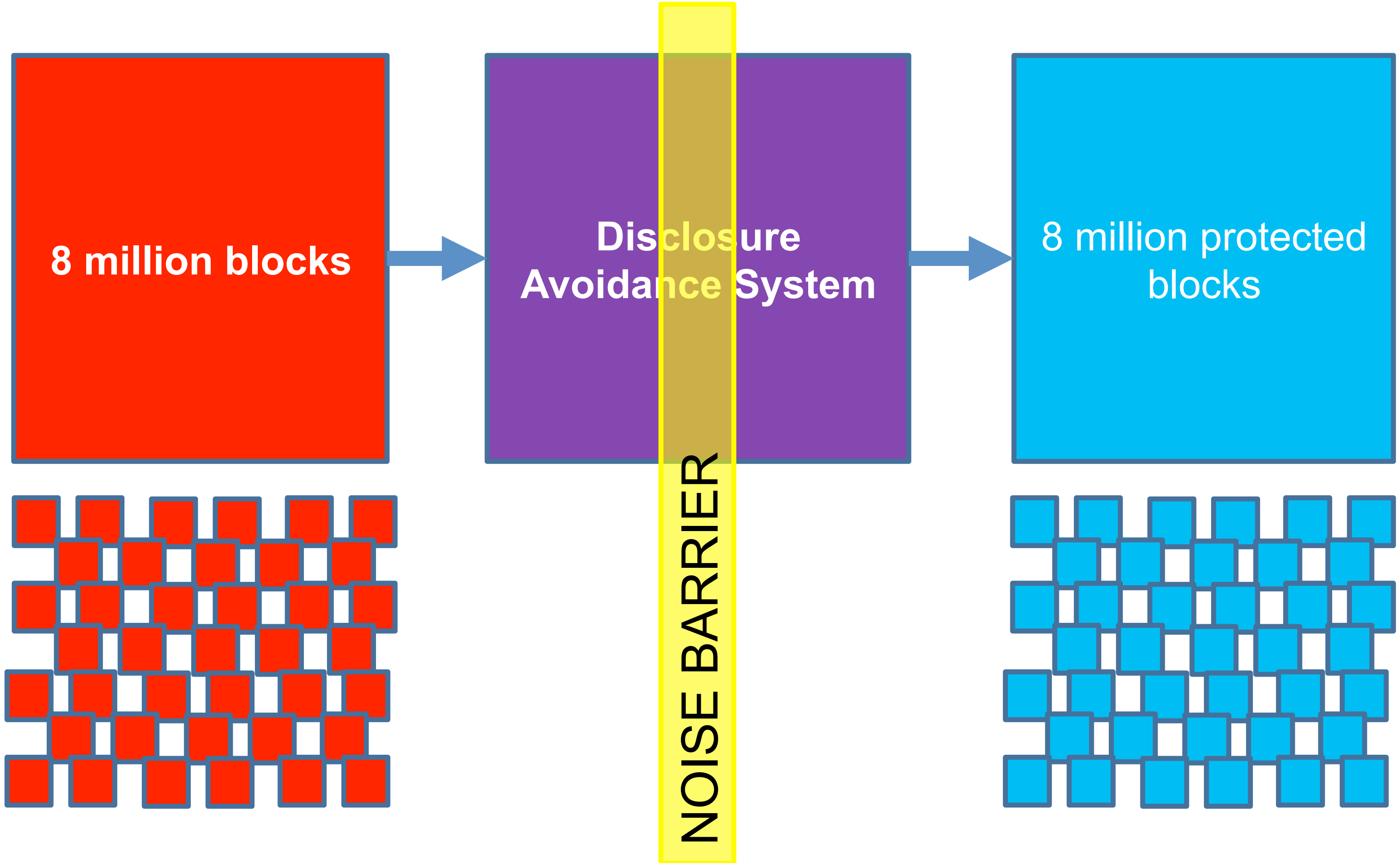
- Treat the *entire census* as a set of queries on histograms.
- Select the specific queries to measure
  - Six *geolevels* (nation, state, county, tract, block group, block)
  - Thousands of queries per *geounit*
  - Billions of queries overall
  - Histogram has billions of cells

Protecting the data



First effort: The block-by-block algorithm

Independently protect each block (parallel composition)



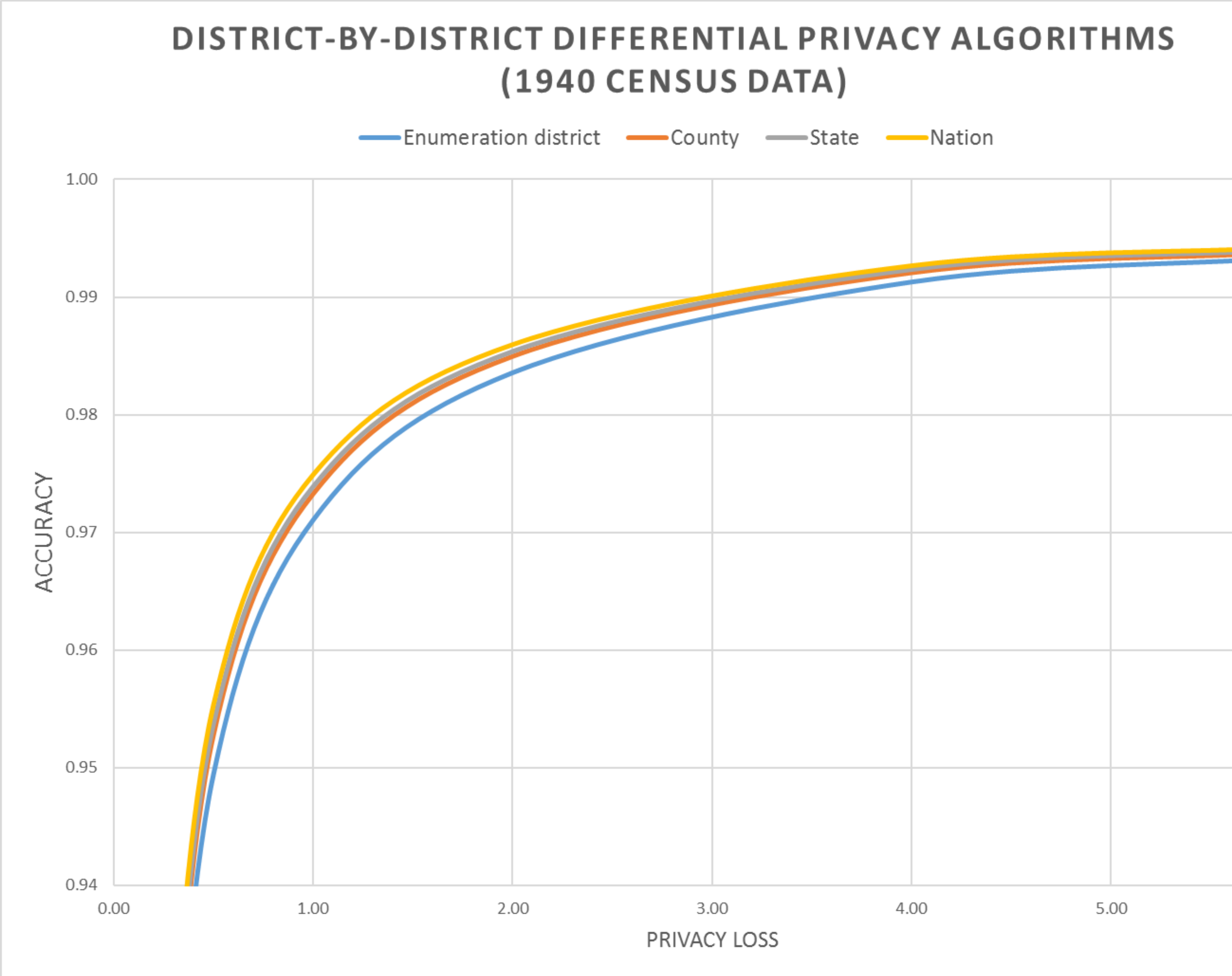
Measure queries for each block; privatize queries; convert results back to microdata

Tested with data from 1940

1940 hierarchy:

- Nation
- State
- County
- Enumeration District

Download from usa.ipums.org



Block-by-block algorithm (also called bottomUp)

Mechanism:

- Select, Measure, Reconstruct separately on each block

Advantages:

- Simple and easy to parallelize
- Privacy cost does not depend on # of blocks
- Releasing DP for one block has same cost as releasing for all

Disadvantages

- Significant error at higher level
- Error adds up
- Variance of each geounit is proportional to the number of blocks it contains

New algorithm: the top-down mechanism

Step 1: Generate national histogram without geographic identifiers.

Step 2: Allocate counts in histogram to each geography "top down."

- National-level measurements - ε_{nat}
- State-level histograms - ε_{state}
- County-level histograms - ε_{county}
- Tract-level histograms - ε_{tract}
- Block-group level histograms - $\varepsilon_{blockgroup}$
- Block-level histograms - ε_{block}

$$\varepsilon = \varepsilon_{nat} + \varepsilon_{state} + \varepsilon_{county} + \varepsilon_{tract} + \varepsilon_{blockgroup} + \varepsilon$$

The top-down algorithm…

Using ε_{state} generate state histograms:

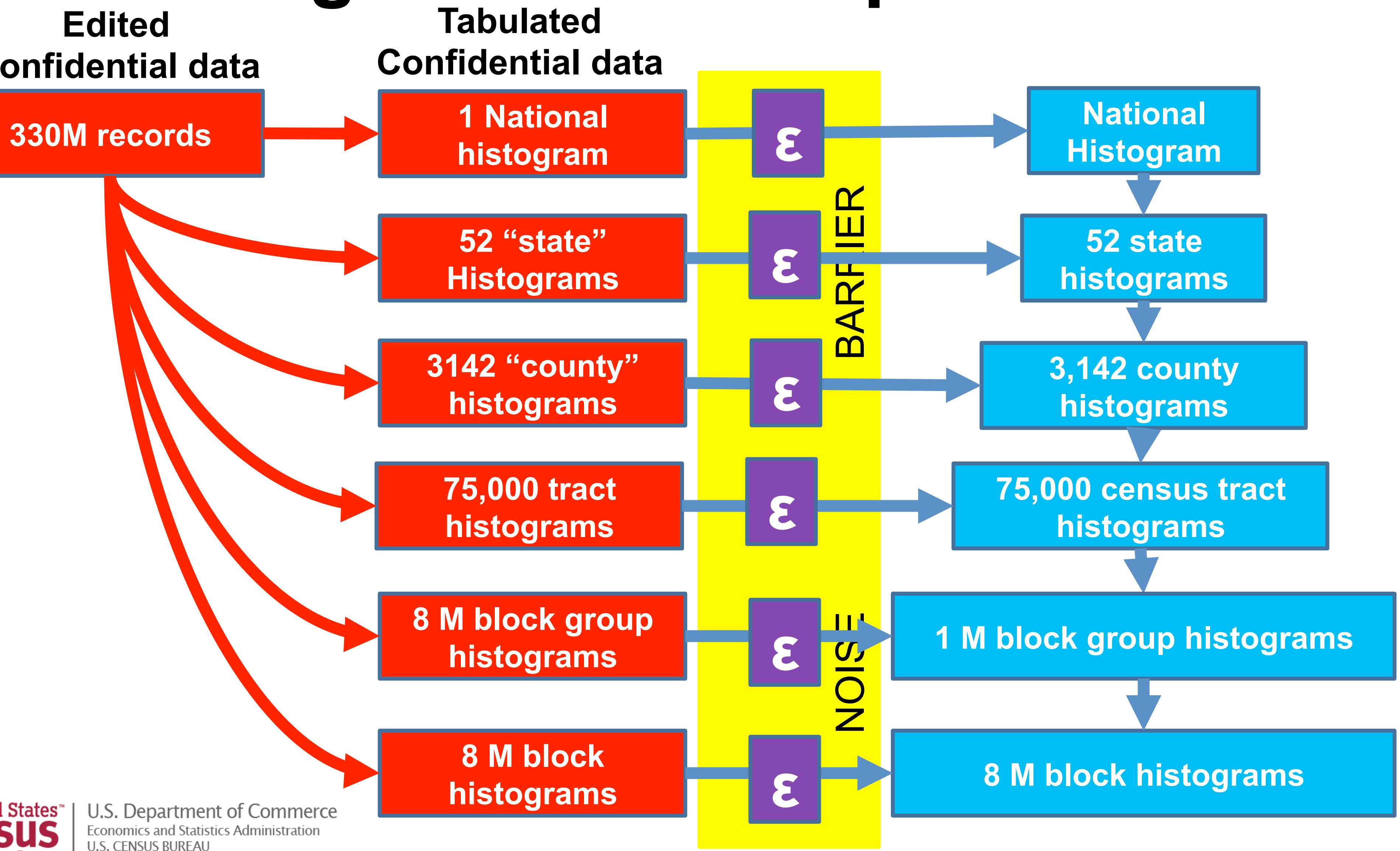
$$H_{AK}^{1},\ H_{AL}^{1},\ldots\ H_{WY}^{1}$$

All histograms are consistent:

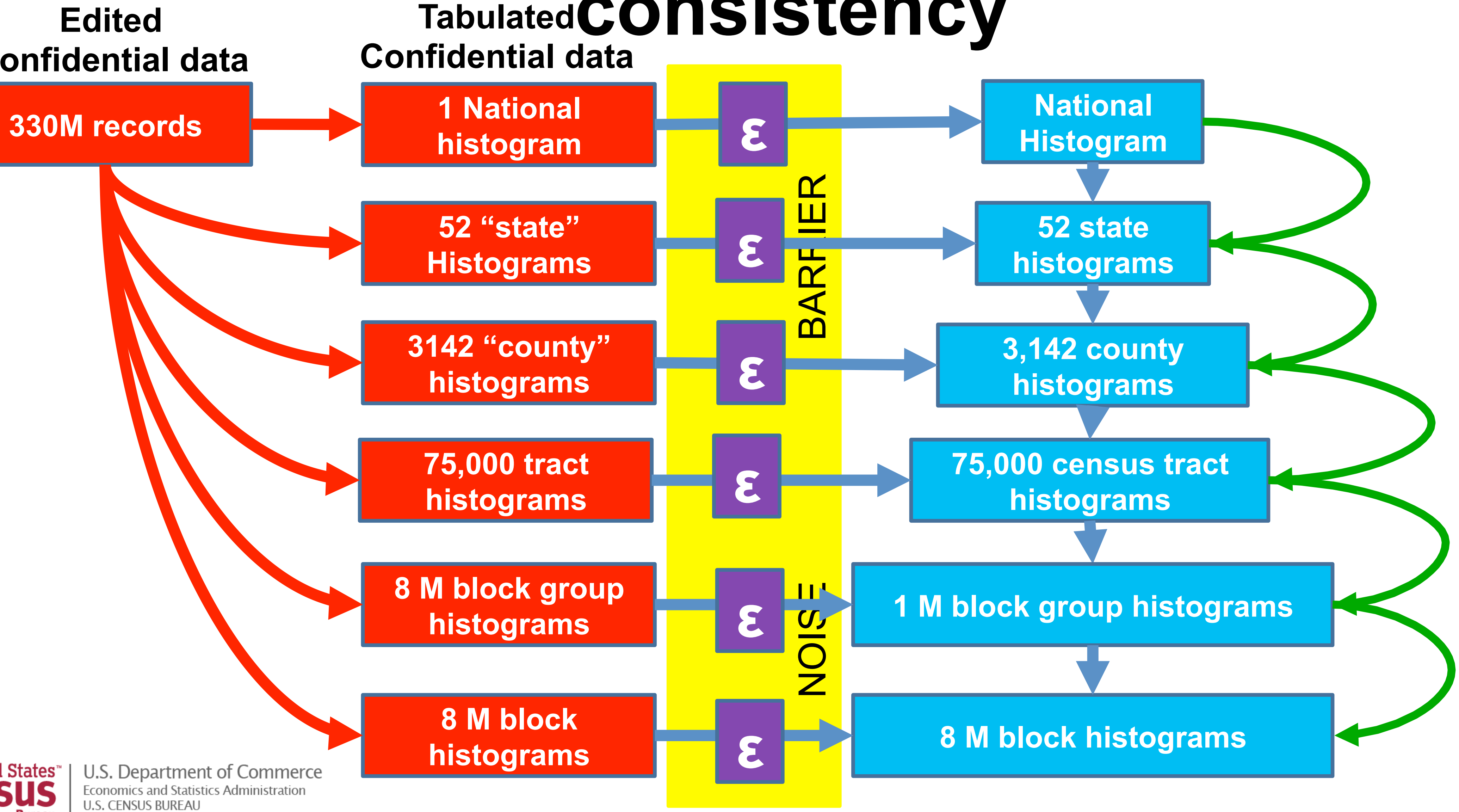
$$\sum_{s \in states} H_{s}^{1} = H^{0}$$

Repeat for each geolevel

New algorithm: the top-down mechanism



Post-process for non-negativity and consistency



op-down framework: alternative view

lational histogram equivalent to table of records:

Age	Race	Sex	Ethnicity	HHGQ

xtend to state-level histograms:

Age	Race	Sex	Ethnicity	HHGQ	State

dd county:

Age	Race	Sex	Ethnicity	HHGQ	State	County

hen add tract, block group, block

Top-Down Framework

Advantages:

- Easy to parallelize
- Each geo-unit can have its own strategy selection
- We use High Dimensional Matrix Mechanism [MMHM18]
- Parallel composition at each geo-level
- Reduced variance for many aggregate regions
- Sparsity discovery
  - *e.g. very few 100+ aged people who combine 5 races*
  - *Once to—down decide a region has no such records in county A, no subregion will have them.*

Post-processing

Each distribution involves (at least) two runs of an optimizer

L_2 solve:

- Generates nonnegative fractional demographics histogram
- State histograms must add up to National histogram (etc.)

L_1 solve:

- Converts fractional histogram to non-negative integer histogram
- Maintain consistency: child histograms must add up to parent
- Integer solutions are fast to find

Evaluating the algorithm

We released runs of the top-down algorithm on data from the 1940 Census.

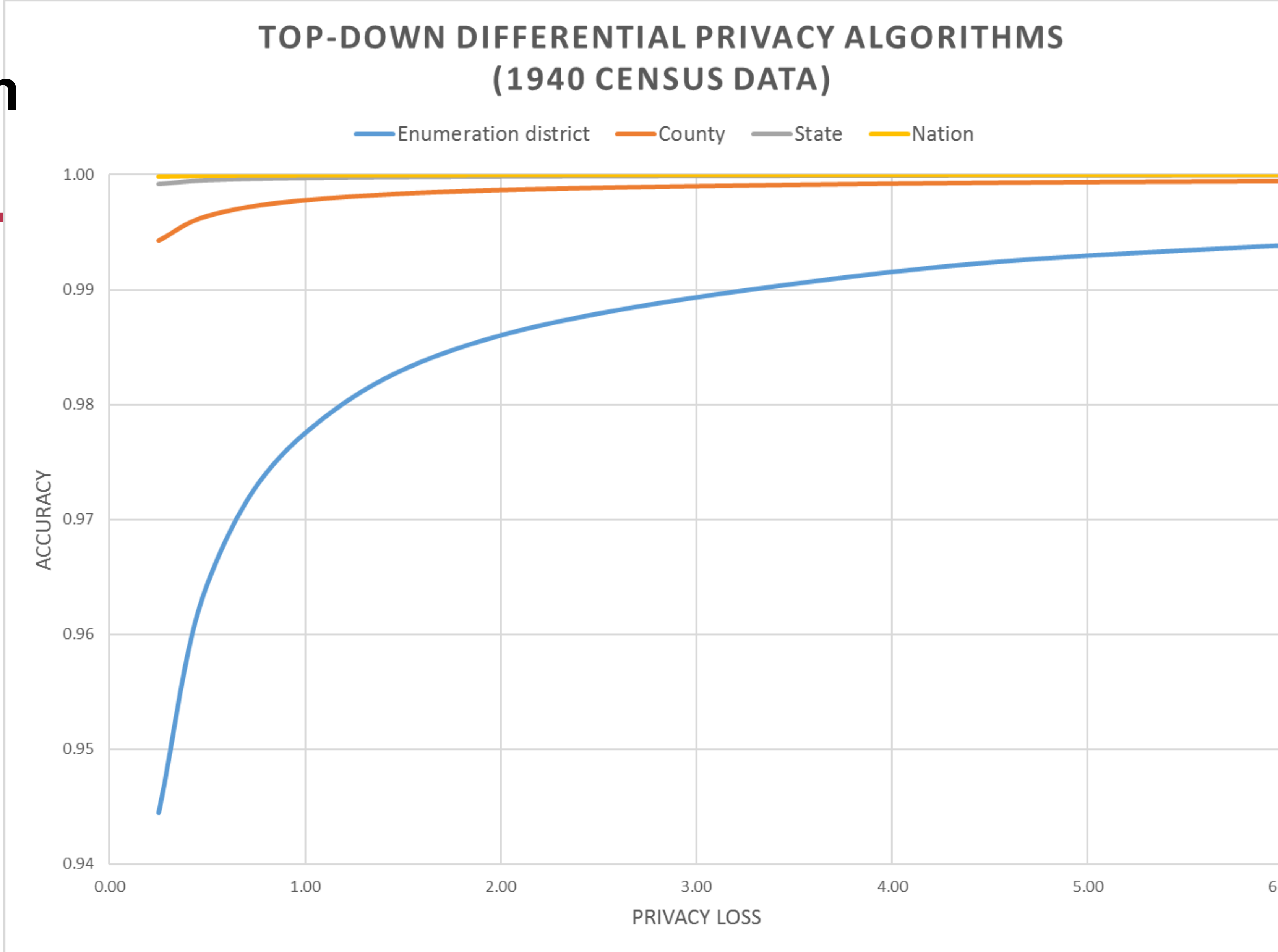
- Epsilon values 0.25 .. 8.0
- Multiple runs at each value of epsilon.

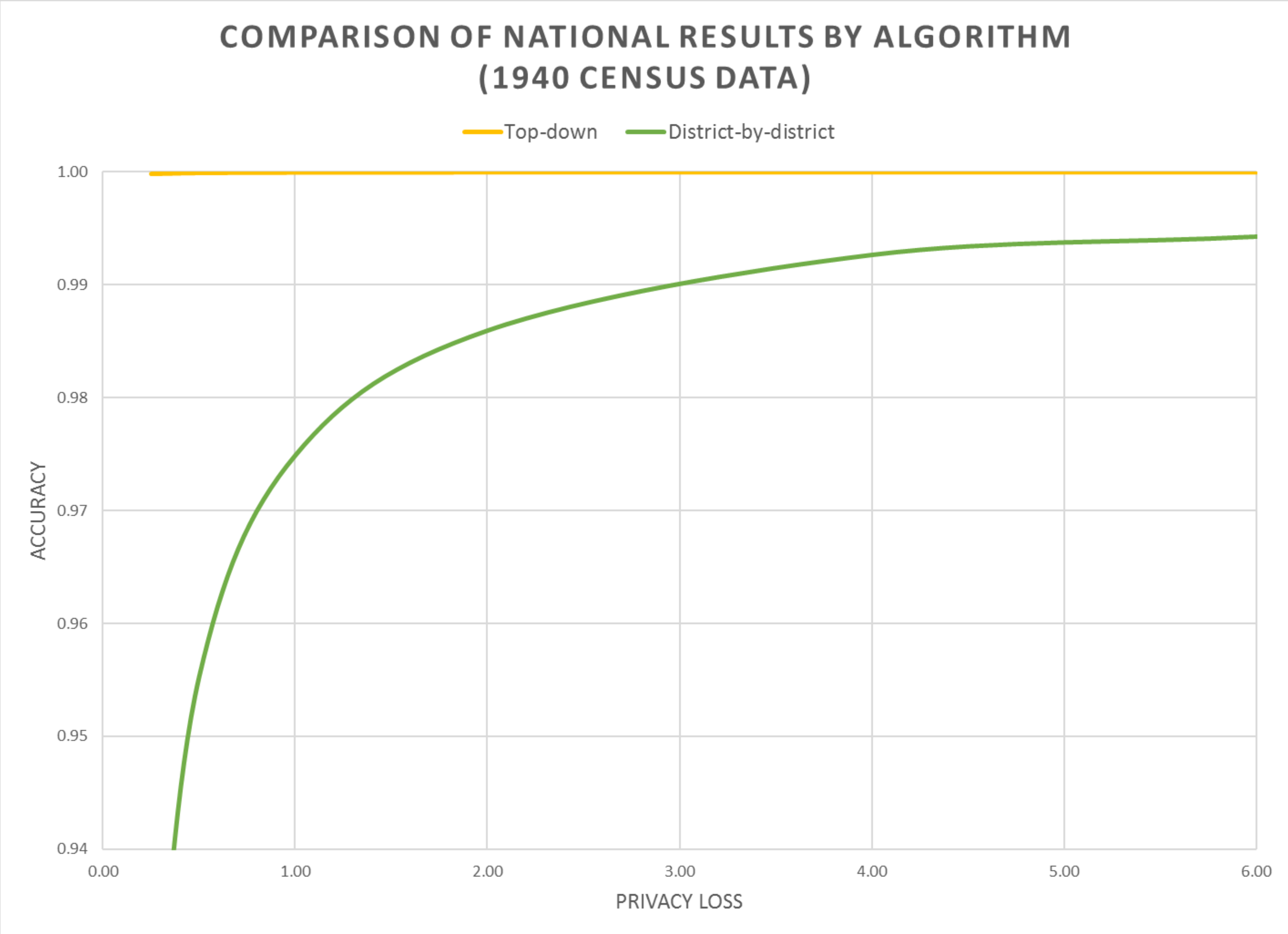
Caveats:

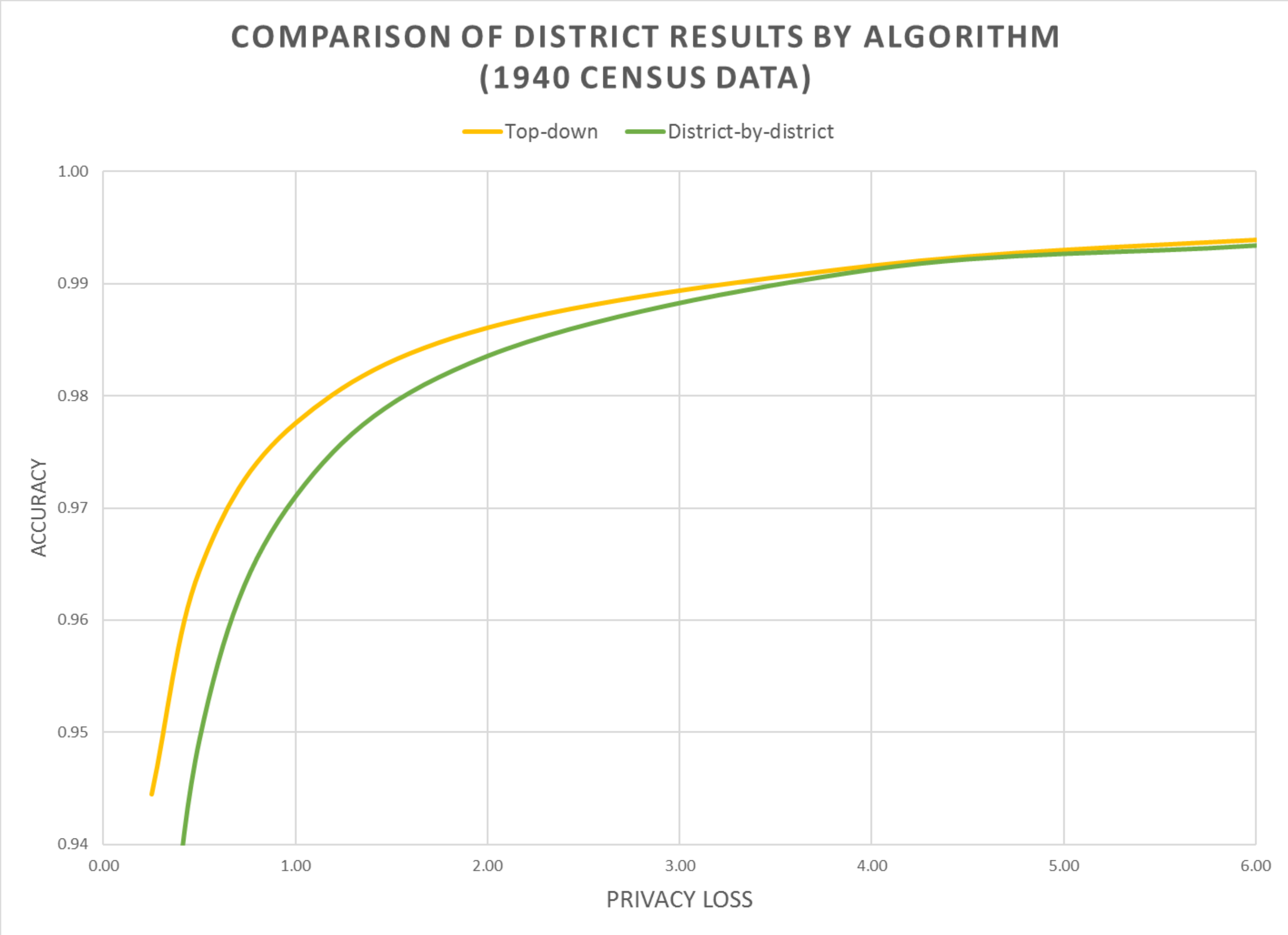
- 1940 data had 4 geography levels: Nation, State, County, Enumeration District.
- 2020 data has 6 levels: Nation, State, County, Tract, Block Group and Block.

- 1940 data has 6 races races / 2020 data has 63 race combinations
- 1940 data has no citizenship (Citizen or non-Citizen)

Top-Down: much more accurate!




COMPARISON OF NATIONAL RESULTS BY ALGORITHM
(1940 CENSUS DATA)
Top-down
District-by-district
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS


COMPARISON OF DISTRICT RESULTS BY ALGORITHM
(1940 CENSUS DATA)
Top-down
District-by-district
ACCURACY
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS

https://dpwiki.org/demo-top-down/dp-demo/sim-top-down/demo.html

Confidential database:

Tiny County pop: 642 f: 318 m: 324					
Ruralland pop: 21 f: 9 m: 12			Urbanville pop: 621 f: 309 m: 312		
RBlock pop: 3	RBlock pop: 7	RBlock pop: 11	UBlock pop: 203	UBlock pop: 207	UBlock pop: 211
f: 1	f: 3	f: 5	f: 101	f: 103	f: 105
m: 2	m: 4	m: 6	m: 102	m: 104	m: 106

Noise Barrier

privatize!

ε 1.0

Published official tabulations:

Tiny County pop: 642 f: 319 m: 323					
Ruralland pop: 21 f: 10 m: 11			Urbanville pop: 621 f: 309 m: 312		
RBlock pop: 8	RBlock pop: 2	RBlock pop: 11	UBlock pop: 199	UBlock pop: 212	UBlock pop: 210
f: 4	f: 2	f: 4	f: 98	f: 107	f: 104
m: 4	m: 0	m: 7	m: 101	m: 105	m: 106

Each rectangle shows the population statistics for a different geographical area. The top is the total population (pop), followed by the number of females (f) and the number of males (m).

ε specifies the privacy loss budget. Click **privatize!** to re-run the privacy mechanism with a different set of random noises.

Try changng the number of females or males that was counted on a block and see how it changes the official tabulations. Or choose one of the sample scenarios listed below.

select	scenario
	balanced rural and urban blocks
	One rural block with a LOT of males.
	One urban block with a LOT of females.

Note: The simulator uses hypothetical (fake) data provided by the user.

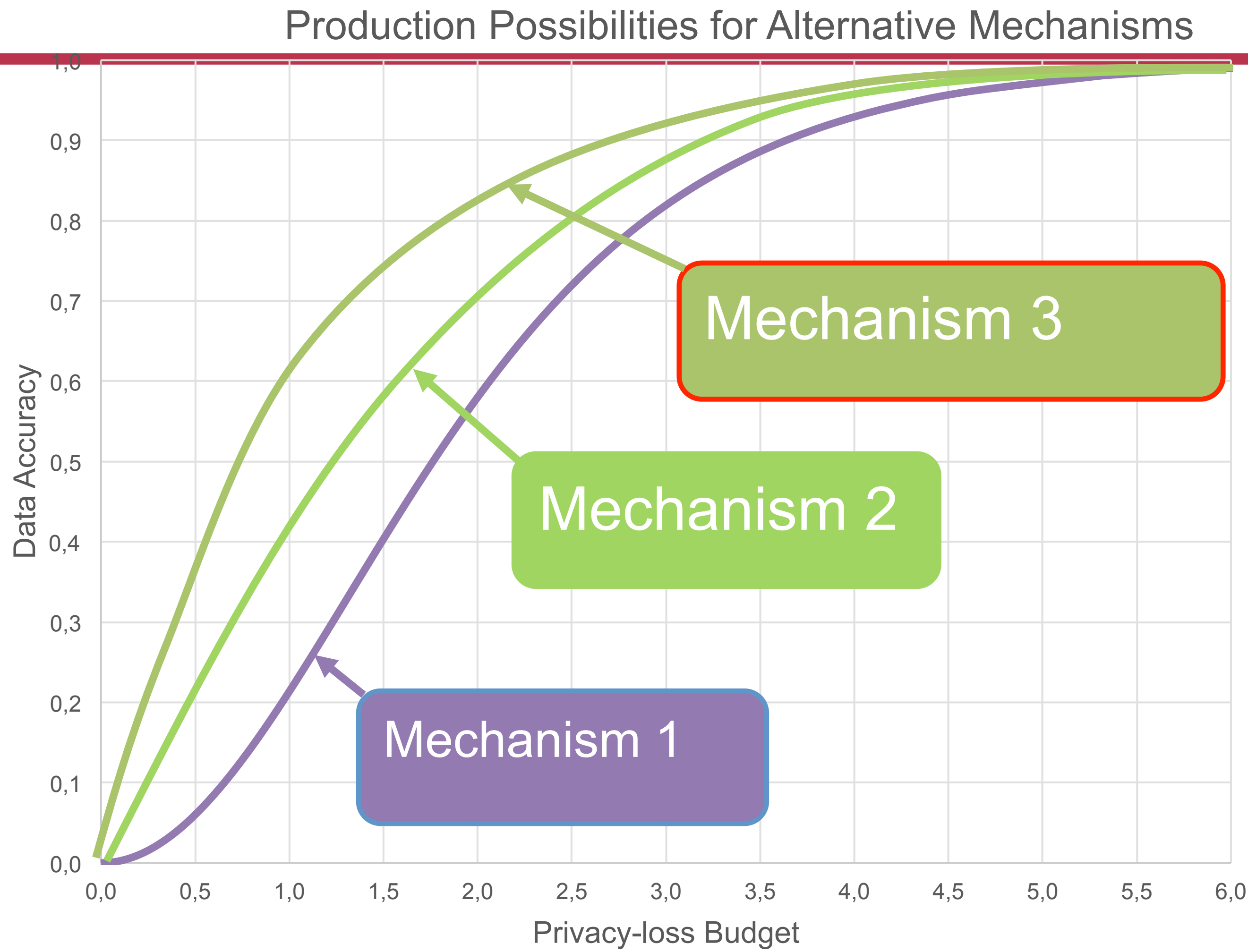
Two public policy choices:

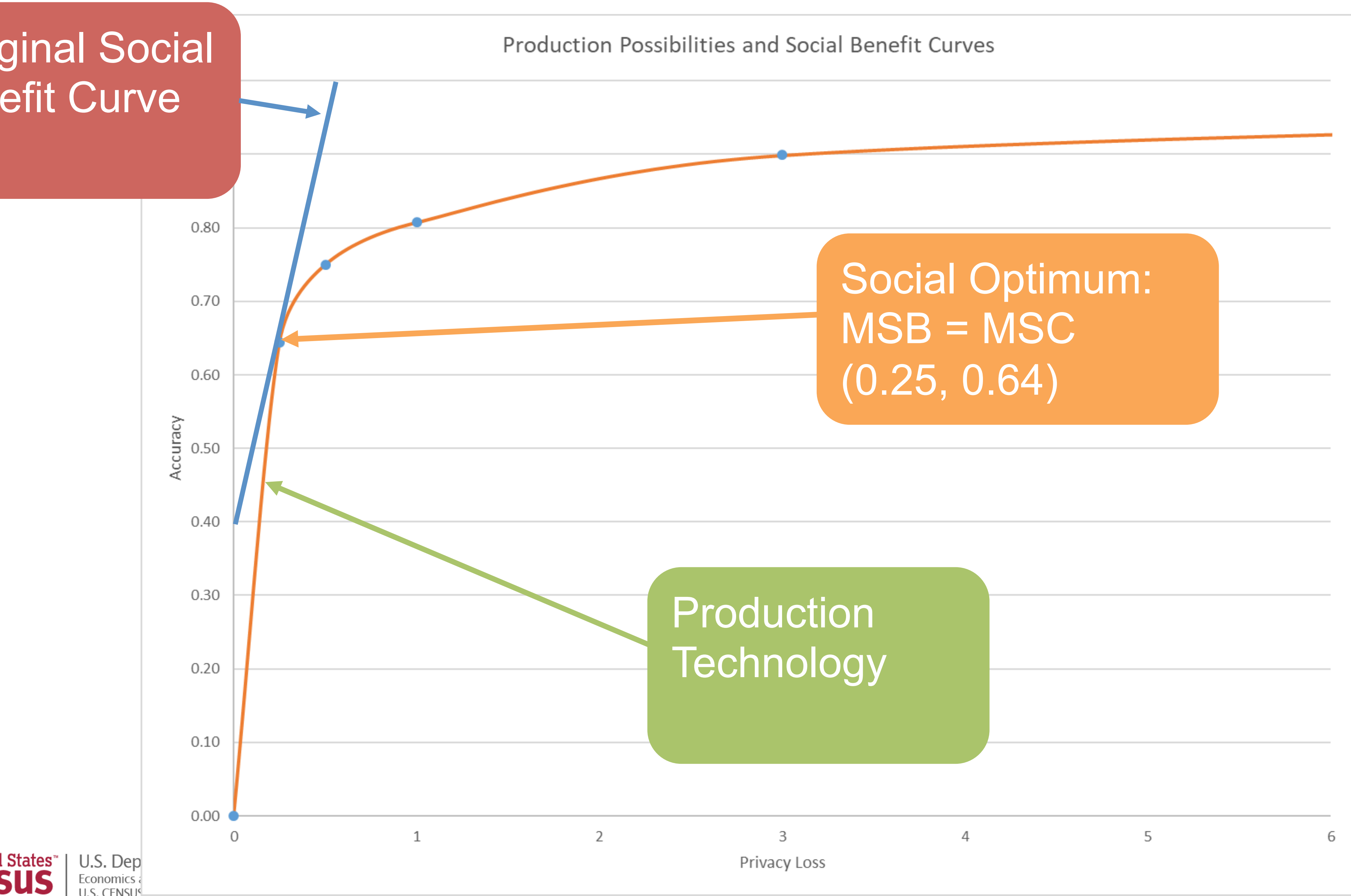
What is the correct value of epsilon?

Where should the accuracy be allocated?

Managing the Tradeoff

Policy Issue: Setting Epsilon





ginal Social
efit Curve
Production Possibilities and Social Benefit Curves
Social Optimum:
MSB = MSC
(0.25, 0.64)
Production
Technology
Accuracy
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
Privacy Loss
0
1
2
3
4
5
6

Policy Decisions: Setting privacy loss budget (ε)

- Global privacy loss budget
- Geographic levels
  - Fraction of **ε** allocated to each level
- Tables
  - Fraction of **ε** allocated to each table or relationship

Policy Issues for the 2020 Census: Invariants

For the 2018 End-to-End test, policy makers wanted exact counts:

- Number of people on each block
- Number of people on each block of voting age
- Number of residences & group quarters on each block

We implemented invariants before we understood their mathematical impact on differential privacy semantics. We then scaled back to four invariants:

- C1: Total population (invariant at the county level for the 2018 E2E)
- ~~C2: Voting-age population (population age 18 and older)~~ (eliminated for the 2018 E2E)
- C3: Number of housing units (invariant at the block level)
- C4: Number of occupied housing units (invariant at the block level)
- C5: Number of group quarters facilities by group quarters type (invariant at the block level)

Scientific Issues for the 2020 Census: Person-Household Joins

The Census creates two kinds of tables:

- Person tables
- Household tables

We can create P & H today.

We are working on P x H and Detailed P, H

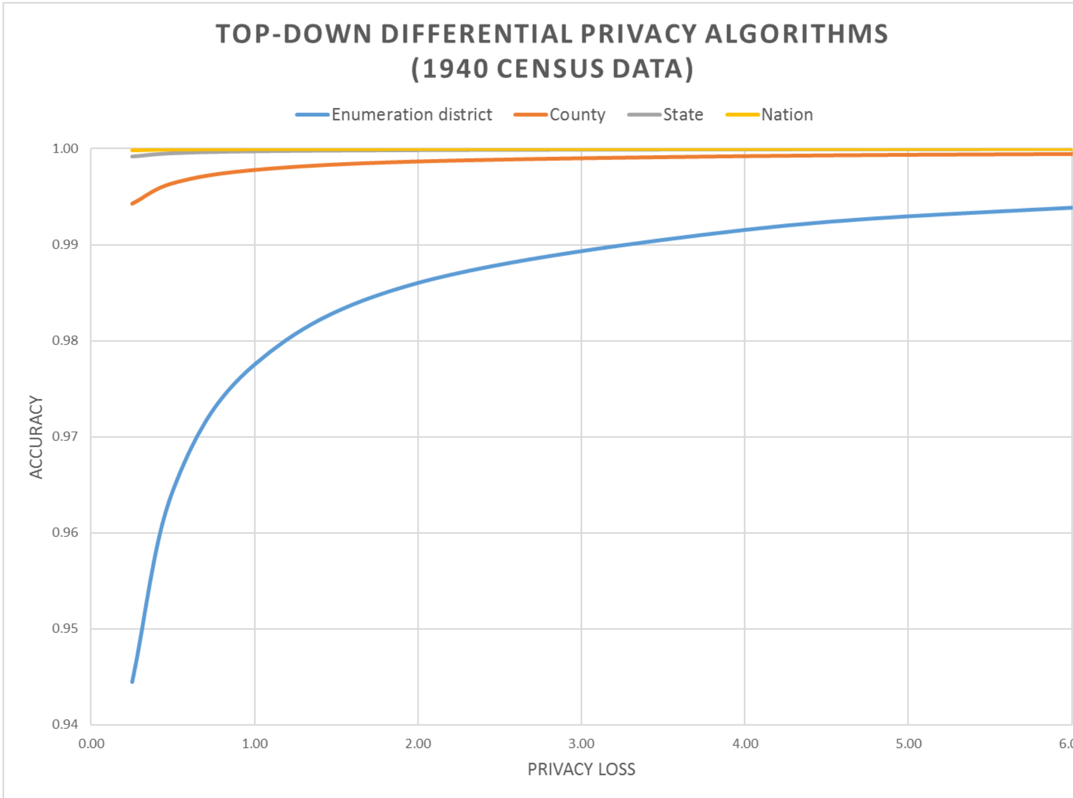
Q(P): # of men living on a block.

Q(H): # of occupied houses on a block.

Q(P x H): # of children in houses headed by a single man.

Scientific Issue for any use of DP: Quality Metrics

What is the measure of "quality" or "utility" in a complex data product?



Options:

- L1 error between "true" data set and "protected" data set
- Impact on an algorithm that uses the data (e.g., redistricting and Voting Rights Act enforcement)

The Choice Problem for Redistricting Tabulations Is More Challenging

In the redistricting application, the fitness-for-use is based on :

- Supreme Court one-person one-vote decision (All legislative districts must have approximately equal populations; there is judicially approved variation)
- *Is statistical disclosure limitation a "statistical method" (permitted by Utah v. Evans) or "sampling" (prohibited by the Census Act, confirmed in Commerce v. House of Representatives)?*
- Voting Rights Act, Section 2: requires majority-minority districts at all levels, when certain criteria are met

The privacy interest is based on:

- Title 13 requirement not to publish exact identifying information
- The public policy implications of uses of detailed race, ethnicity and citizenship

Organizational Challenges

Process documentation

- All uses of confidential data need to be tracked and accounted.

Workload identification

- All desired queries on MDF should be known in advance.
- Required accuracy for various queries should be understood.
- Queries outside of MDF must also be pre-specified

Correctness and Quality control

- Verifying implementation correctness.
- Data quality checks on tables cannot be done by looking at raw data.

Data User Challenges

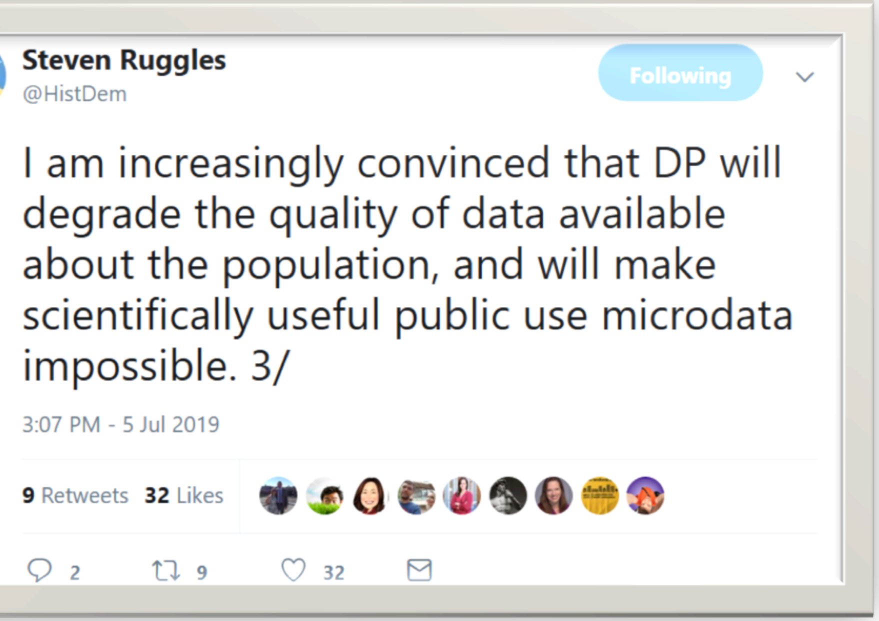
fferential privacy is not widely known or understood.

any data users want highly accurate data reports on small areas.

- Some are anxious about the intentional addition of noise.
- Some are concerned that previous studies done with swapped data might not b replicated if they used DP data.

any data users believe they require access to Public Use Microdata.

sers in 2000 and 2010 didn't know the error introduced by swapping an her protections applied to the tables and PUMS.


Steven Ruggles
@HistDem
Following

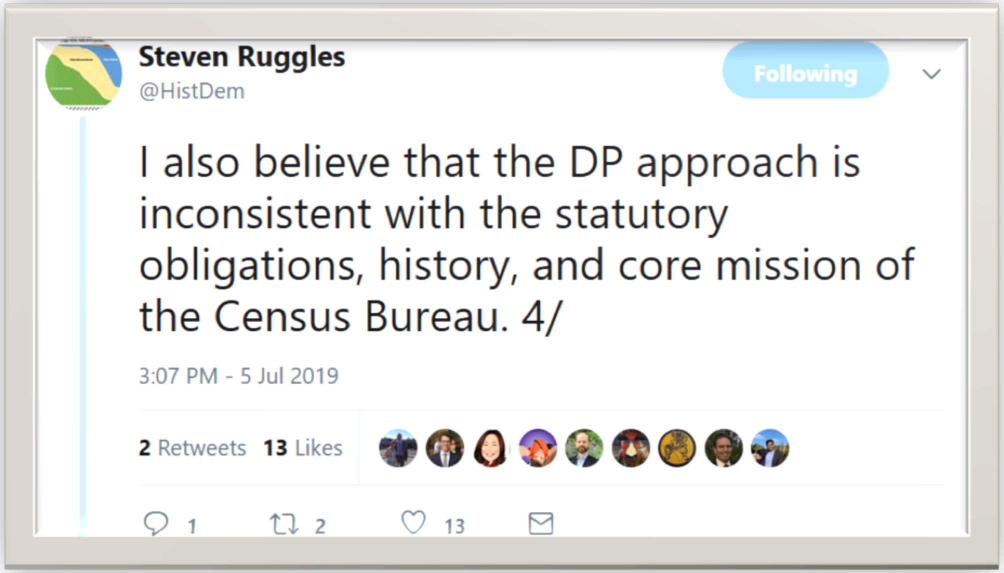
I am increasingly convinced that DP will degrade the quality of data available about the population, and will make scientifically useful public use microdata impossible. 3/

3:07 PM - 5 Jul 2019

9 Retweets 32 Likes

2 9 32

Concerns and Responses


Steven Ruggles
@HistDem
Following

I also believe that the DP approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau. 4/

3:07 PM - 5 Jul 2019

2 Retweets 13 Likes

1 2 13

Redistricting and Exact Counts

- n the US, legislative districts must have equal size.
  - Decennial Census counts of each block are the "official counts."
- Some data users are concerned that adding noise to the counts will make them unfit for use.
- However:
  - Evaluation of districts is based on official decennial counts; these data are used for 10 years.
  - Noise added by DP is significantly less than noise added by other statistical methods currently in use

STEVEN RUGGLES

Regents Professor of History and Population Studies
Director, Institute for Social Research and Data Innovation
50 Willey Hall
University of Minnesota
ruggles@umn.edu
(612) 624-5818

Ruggles Concerns

Differential privacy is not a measure of identifiability

Differential privacy does not measure disclosure risk

Differential Privacy is not concerned with re-identification of respondents

- “DP prohibits revealing *characteristics* of an individual even if the *identity* of that individual is effectively concealed
- “This is a radical departure from established census law and precedent
- “The Census Bureau has been disseminating individual-level *characteristics* routinely since the first microdata in 1962

Organized attack on the move to differential privacy

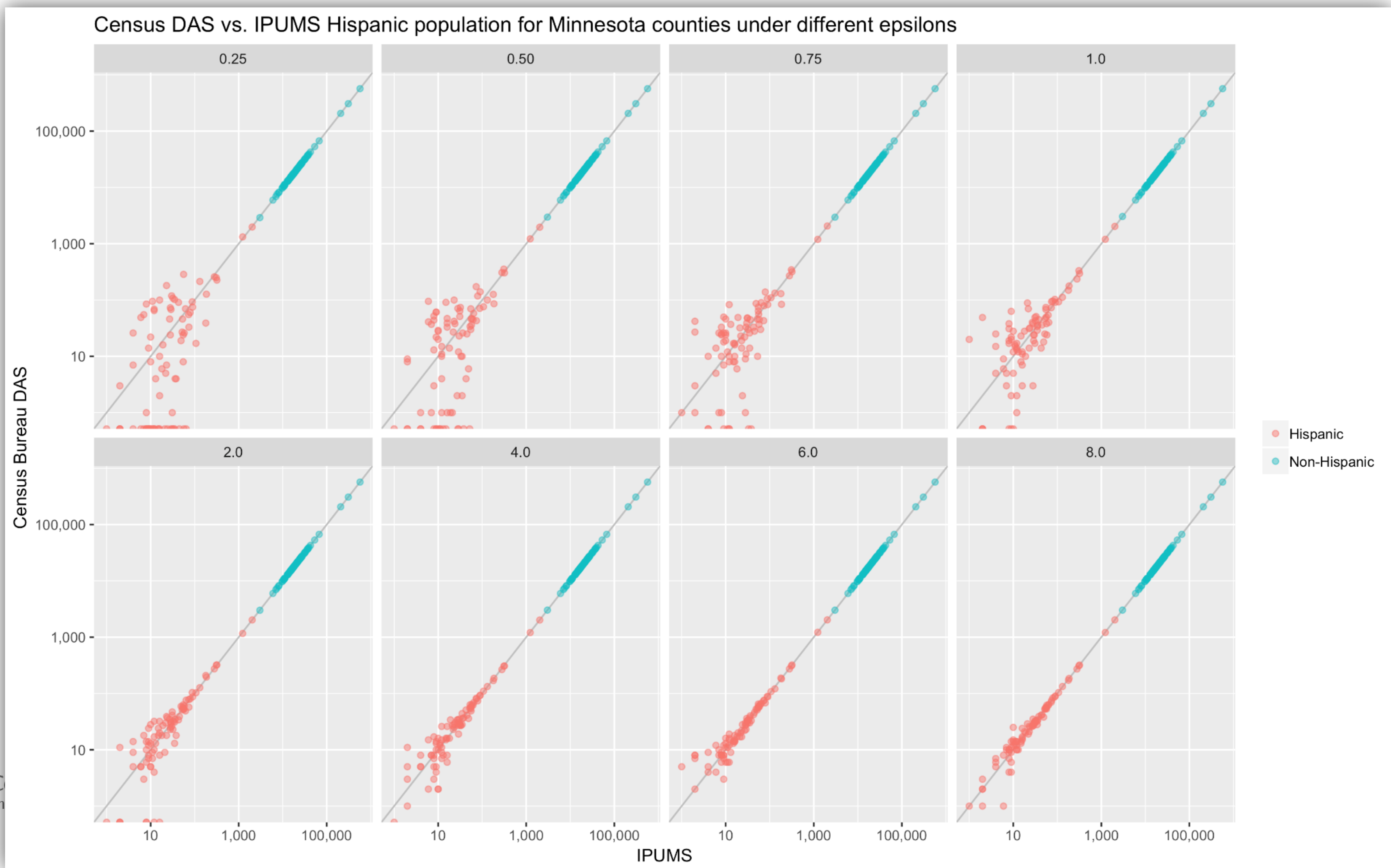
STEVEN RUGGLES

Regents Professor of History and Population Studies
Director, Institute for Social Research and Data Innovation
50 Willey Hall
University of Minnesota
ruggles@umn.edu
(612) 624-5818

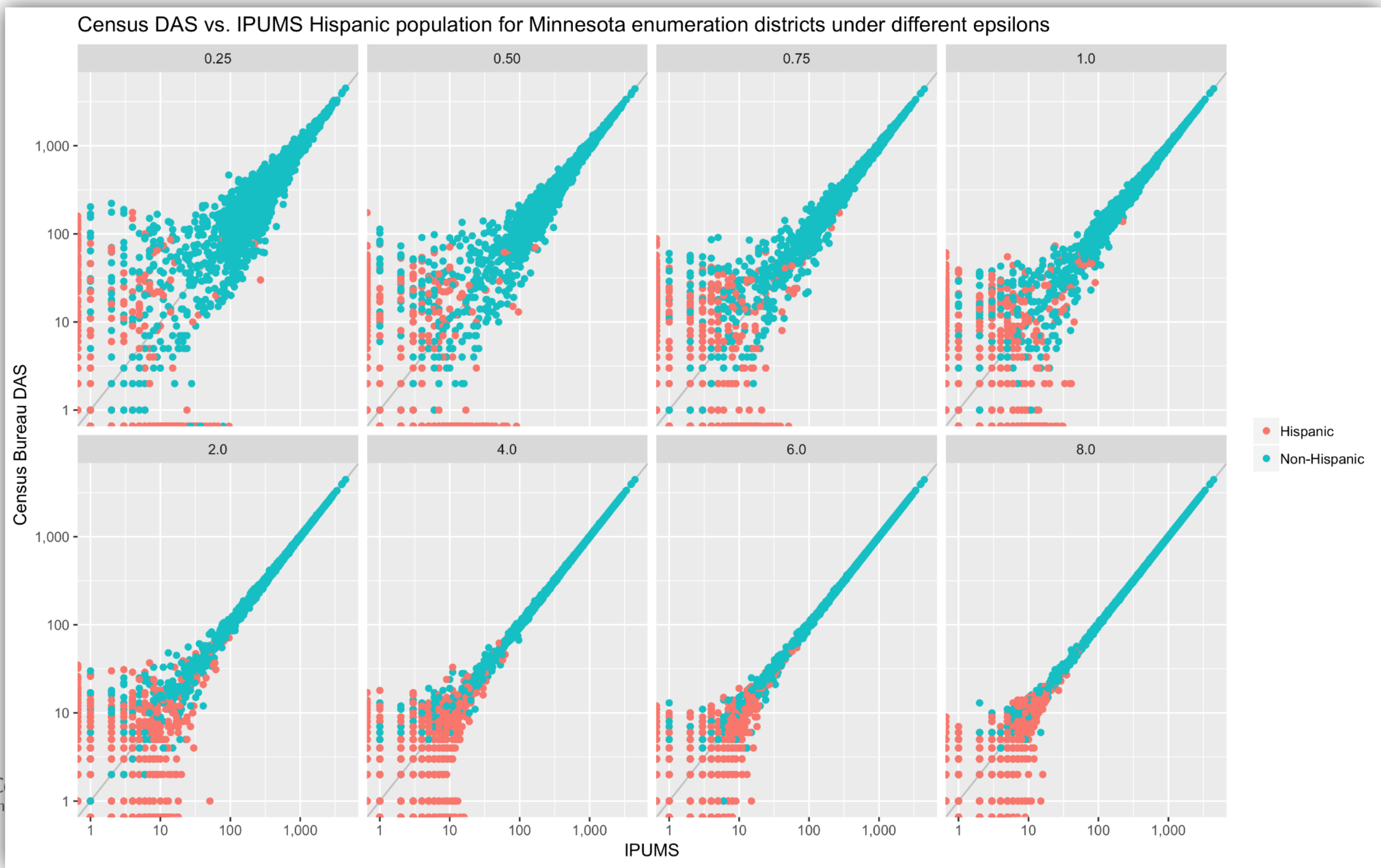
Concerns:

- “Differential privacy will degrade the quality of data available about the population, and will probably make scientifically useful public use microdata impossible
- The differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau”

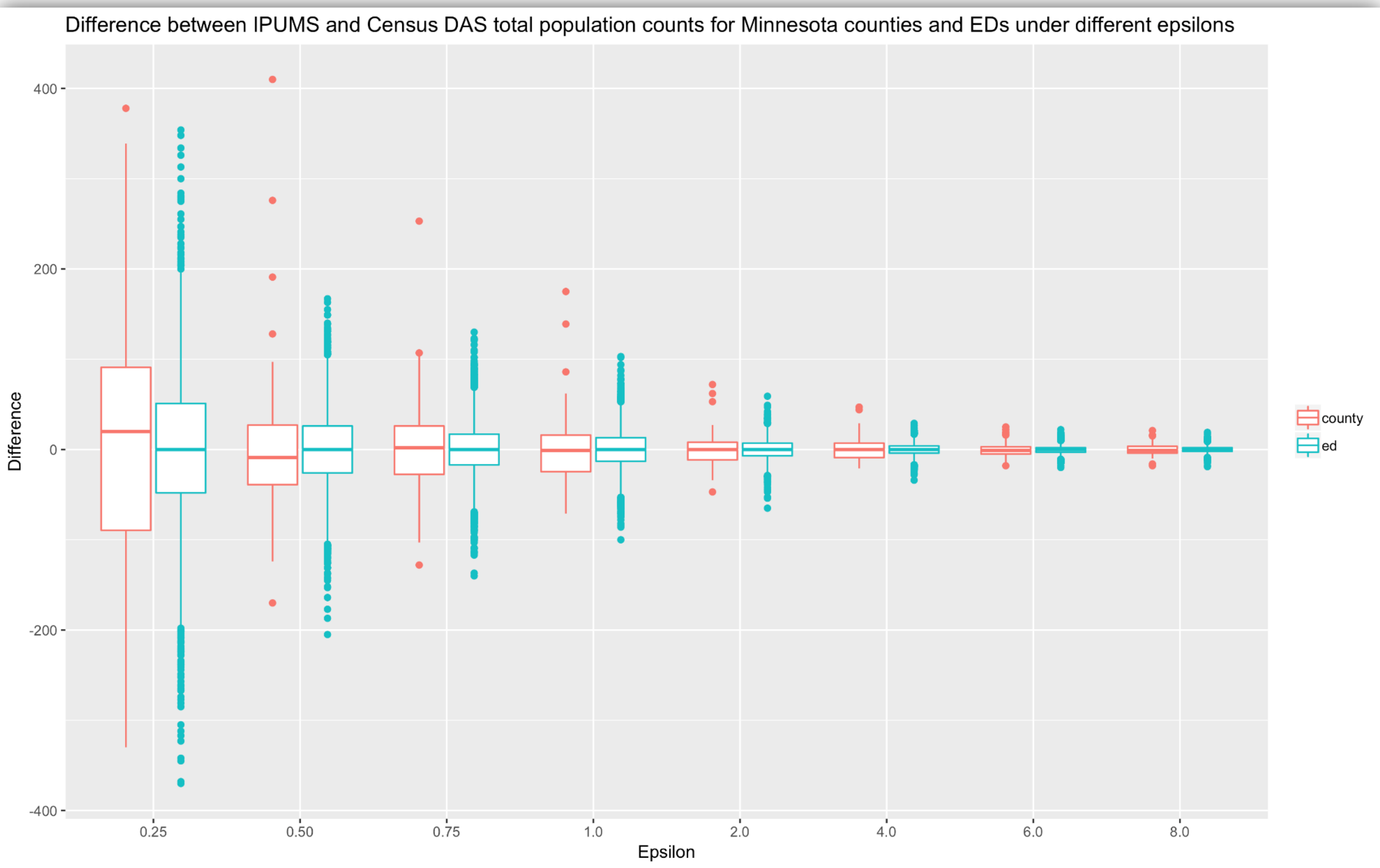
Analysis of population variances
David Van Riper & Tracy Kugler, IPUMS (APDU 2019)



Analysis of population variances
David Van Riper & Tracy Kugler, IPUMS (APDU 2019)



Analysis of population variances
David Van Riper & Tracy Kugler, IPUMS (APDU 2019)



For more information…

practice

DOI:10.1145/3287287

Article development led by acmqueue
queue.acm.org

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to *database reconstruction attacks* (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose of the decennial census, the U.S. Congress has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey responses. All census publications must uphold the confidentiality standard specified by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying "the data furnished by any particular establishment or individual." This section prohibits the Census Bureau from publishing respondents' names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency *accurately* reports there are two persons living on a block and the average age of the block's residents is 35, that would constitute an improper disclosure of personal information, because one of the residents could look up the data, subtract their contribution, and infer the age of the other.

Communications of ACM March 2019
Garfinkel & Abowd

Can a set of equations keep U.S. census data private?
By Jeffrey Mervis
Science
Jan. 4, 2019 , 2:50 PM

http://bit.ly/Science2019C1

More Background on the 2020 Disclosure Avoidance System

September 14, 2017 CSAC (overall design)
https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf

August, 2018 KDD'18 (top-down v. block-by-block)
https://digitalcommons.ilr.cornell.edu/ldi/49/

October, 2018 WPES (implementation issues)
https://arxiv.org/abs/1809.02201

October, 2018 *ACMQueue* (understanding database reconstruction)
https://digitalcommons.ilr.cornell.edu/ldi/50/ or
https://queue.acm.org/detail.cfm?id=3295691

Memorandum 2019.13: Disclosure Avoidance System Design Parameters
https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_13.html