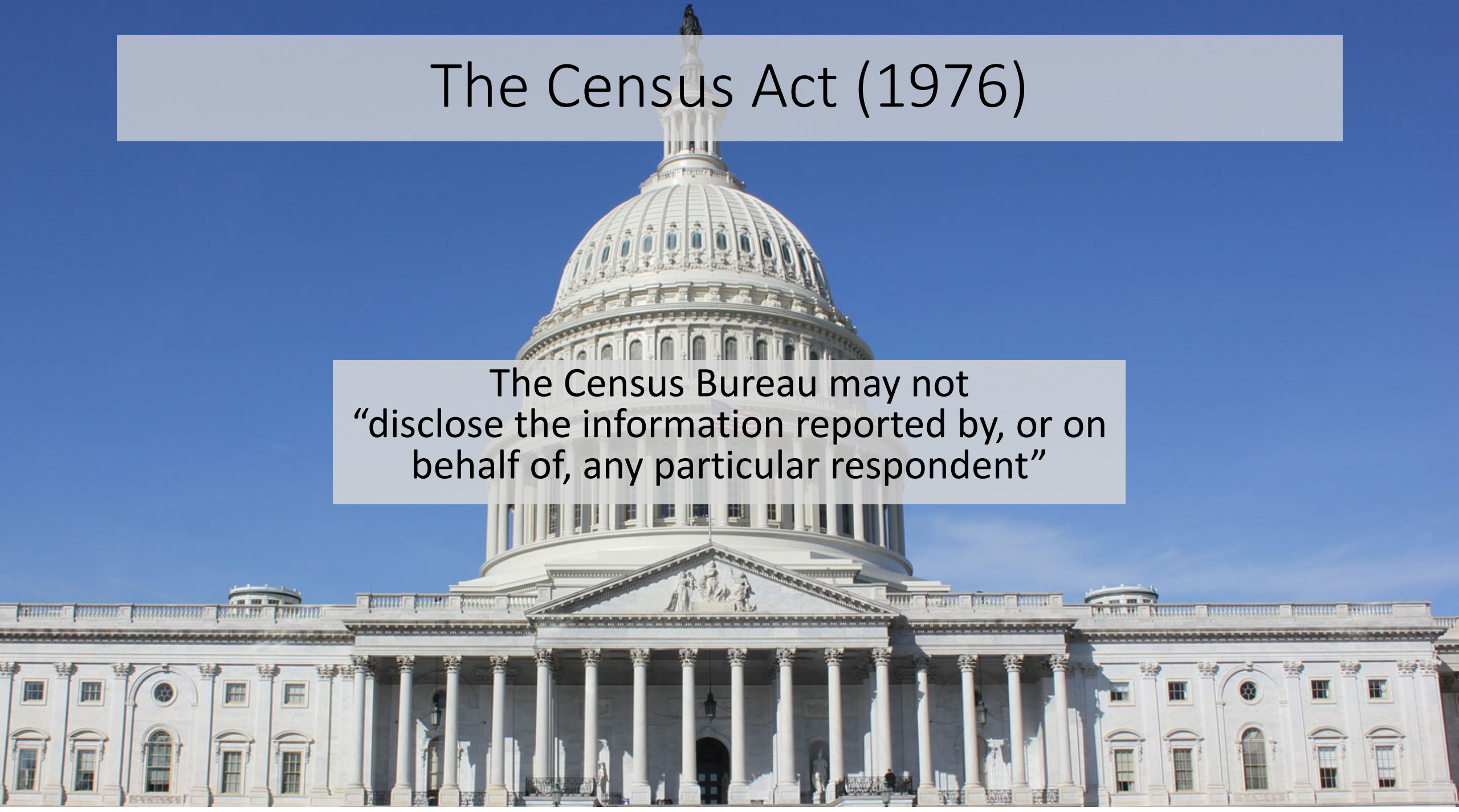
A Kerfuffle Differential Privacy and the 2020 Census

Aloni Cohen
University of Chicago

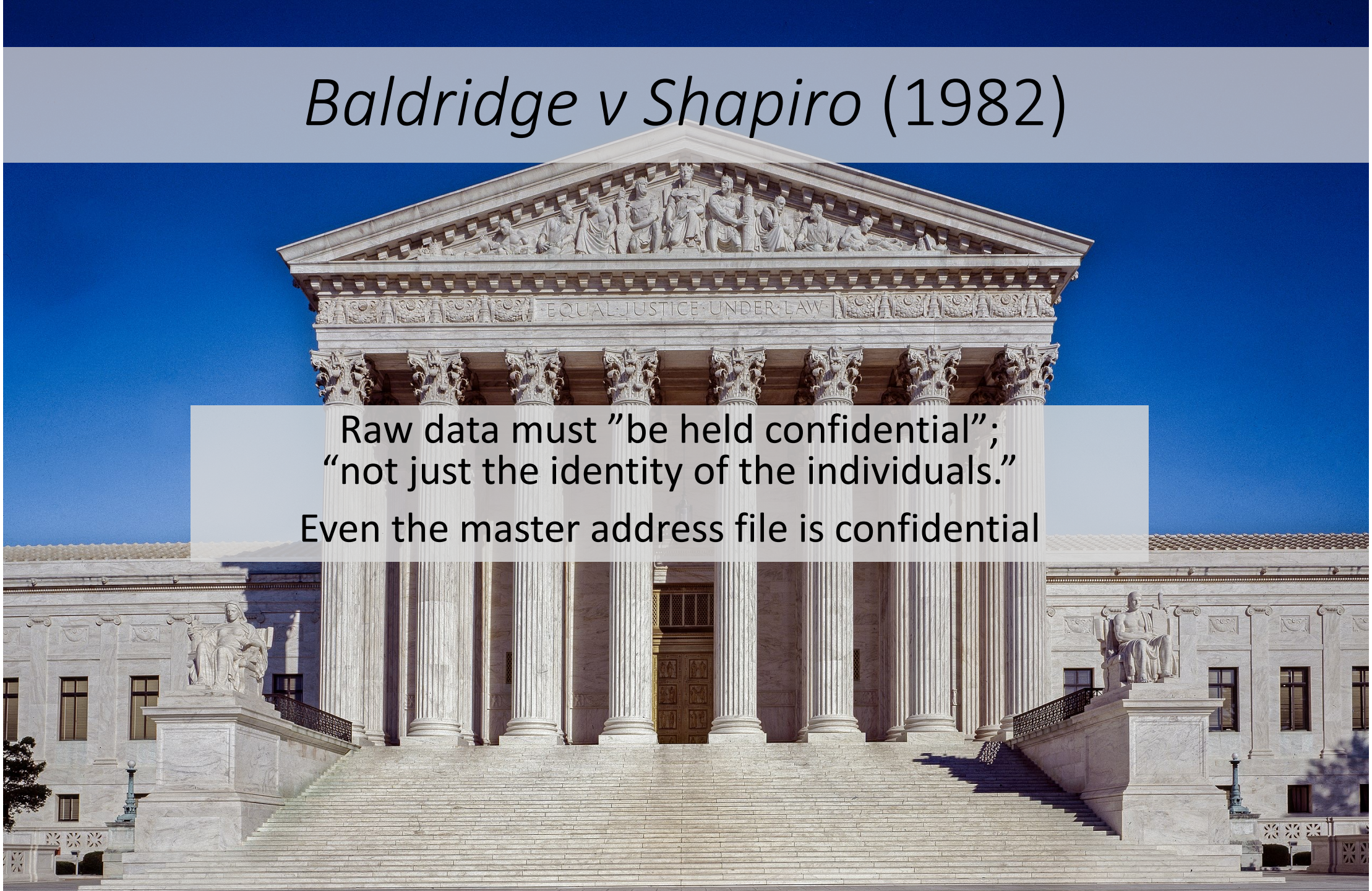
Workshop on Societal Considerations & Applications
Simons Institute
8 Nov 2022

The Census Act (1976)



The Census Bureau may not
"disclose the information reported by, or on behalf of, any particular respondent"

Baldridge v Shapiro (1982)



Raw data must ”be held confidential”;
“not just the identity of the individuals.”

Even the master address file is confidential

Implications of Differential Privacy for Census Bureau Data and Scientific Research

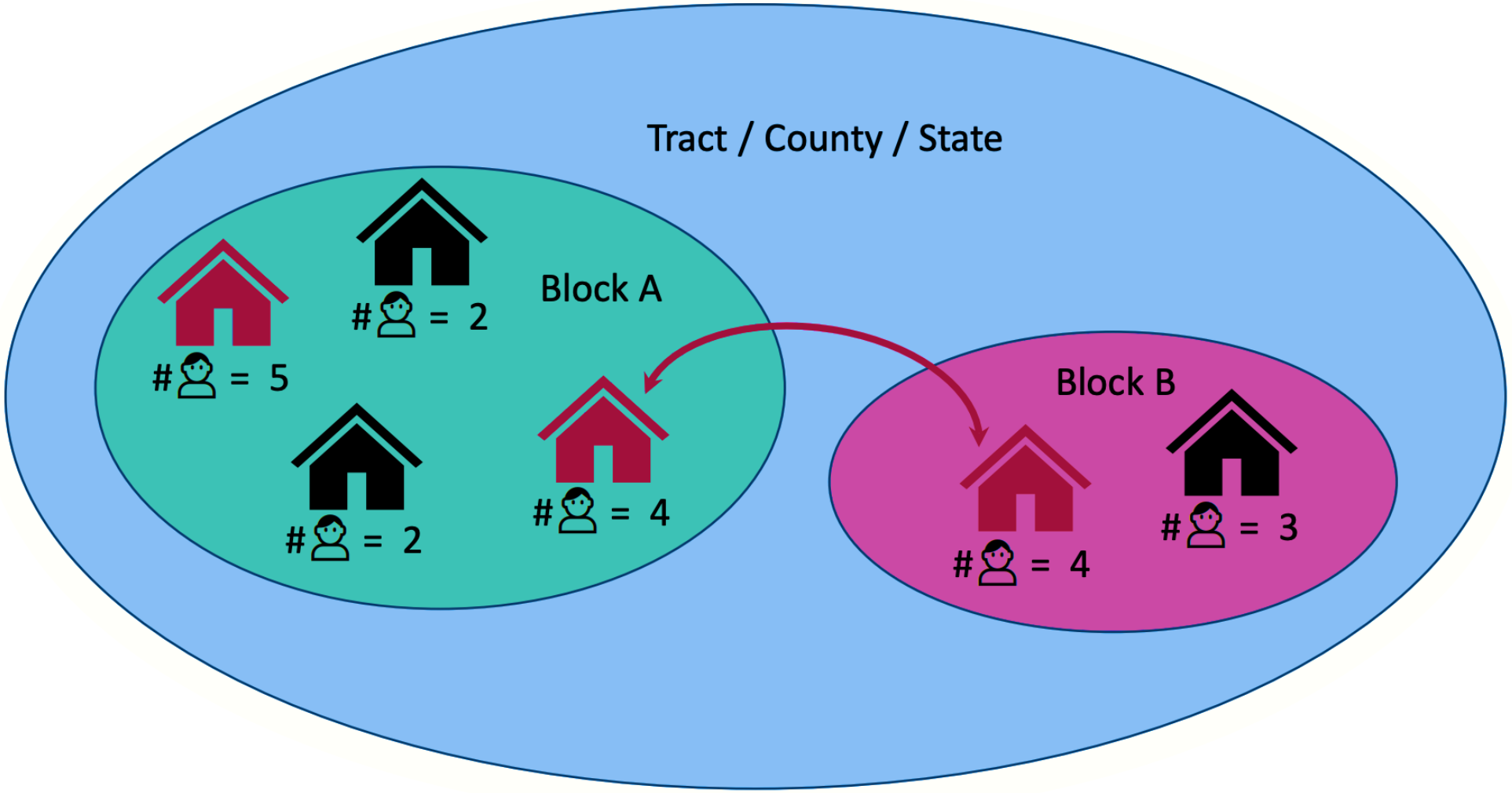
Task Force on Differential Privacy for Census Data†
Institute for Social Research and Data Innovation (ISRDI)
University of Minnesota

December 2018
Version 5.1
Working Paper No. 2018-6
https://doi.org/10.18128/MPC2018-6

This report was prepared by Steven Ruggles (ISRDI) with the assistance of Margo J. Anderson (University of Wisconsin-Milwaukee), Jane Bambauer (Arizona State University), Michael Davern (NORC), Reynolds Farley (University of Michigan),

The meaning of the law is clear and unambiguous: census publications must ensure that the responses of particular identified persons cannot be determined from census publications. To comply with the law, it is not necessary to mask the *characteristics* of individuals; rather, it is necessary to mask the *identity* of individuals. Thus, for the past six decades the Census Bureau disclosure control strategy has focused on targeted strategies to prevent re-identification attacks, so that an outside adversary cannot positively identify which person provided a particular response. The protections in place—sampling, swapping, suppression of geographic information and extreme values, imputation, and perturbation—have worked extremely well to meet this standard. Indeed, there is not a single documented case of anyone outside the Census Bureau revealing the responses of a particular identified person by breaking into public use decennial census or ACS data.

Swapping (1990 – 2010)



2010 Reconstruction

John Abowd @john_abowd · Apr 7, 2019

Replying to @john_abowd

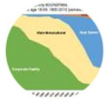
1. First, let's get the facts straight: the U.S. Census Bureau reconstructed 100% of the 2010 Census micro-data records (308,745,538 persons).

2 1 6

John Abowd @john_abowd · Apr 7, 2019

3. The reconstructed records matched the confidential data (2010 CEF) exactly (every single bit) for 46% of the population (142 million people) and allowing age +/- 1 year for 71% of the population (219 million people).

1 5



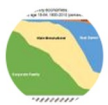
Steven Ruggles @HistDem · Apr 20, 2021

1.I prepared a report for the Plaintiffs in the Alabama v. Department of Commerce lawsuit over differential privacy in the census, available here: users.hist.umn.edu/~ruggles/censi...

UNITED STATES DISTRICT COURT FOR THE MIDDLE DISTRICT OF ALABAMA EASTERN DIVISION

THE STATE OF ALABAMA, *et al.*, Plaintiffs, v. UNITED STATES DEPARTMENT OF COMMERCE; GINA RAIMONDO, *et al.*, Defendants.	CASE NO. 3:21-cv-00211-RAH-ECM-KCN

1 16 17

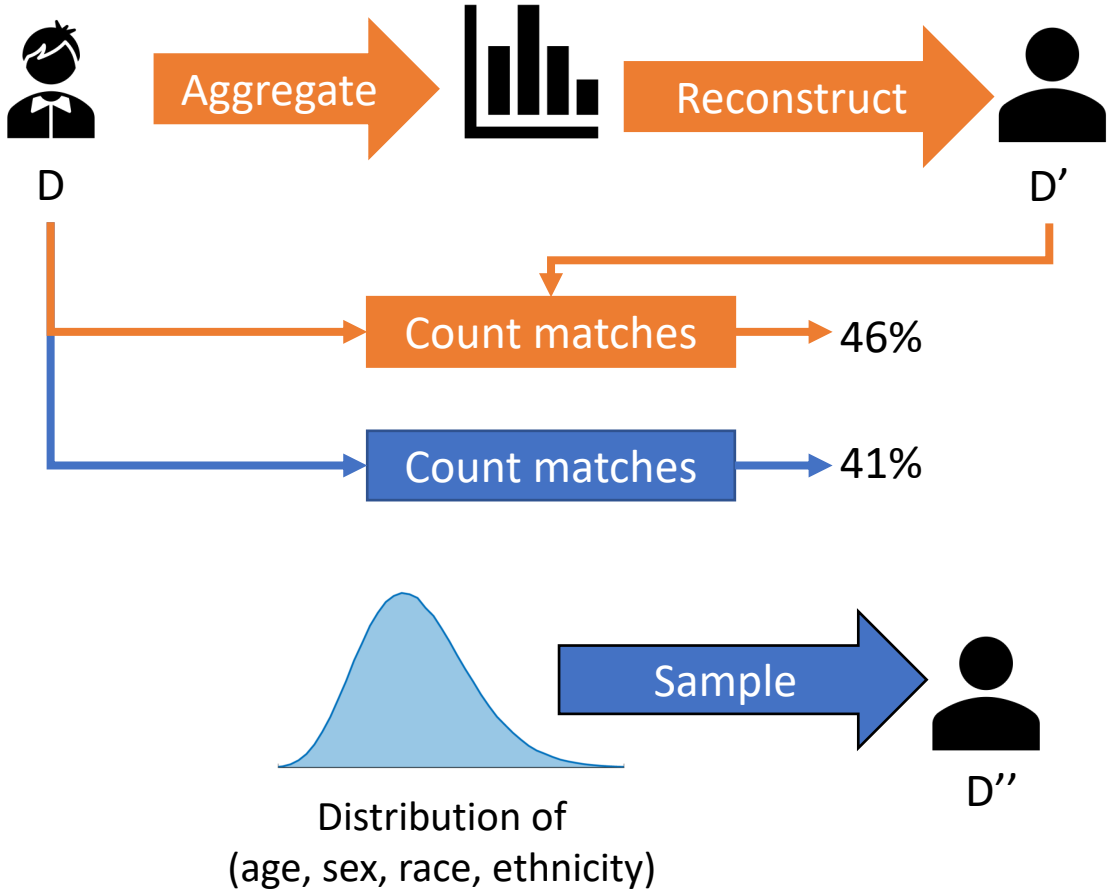


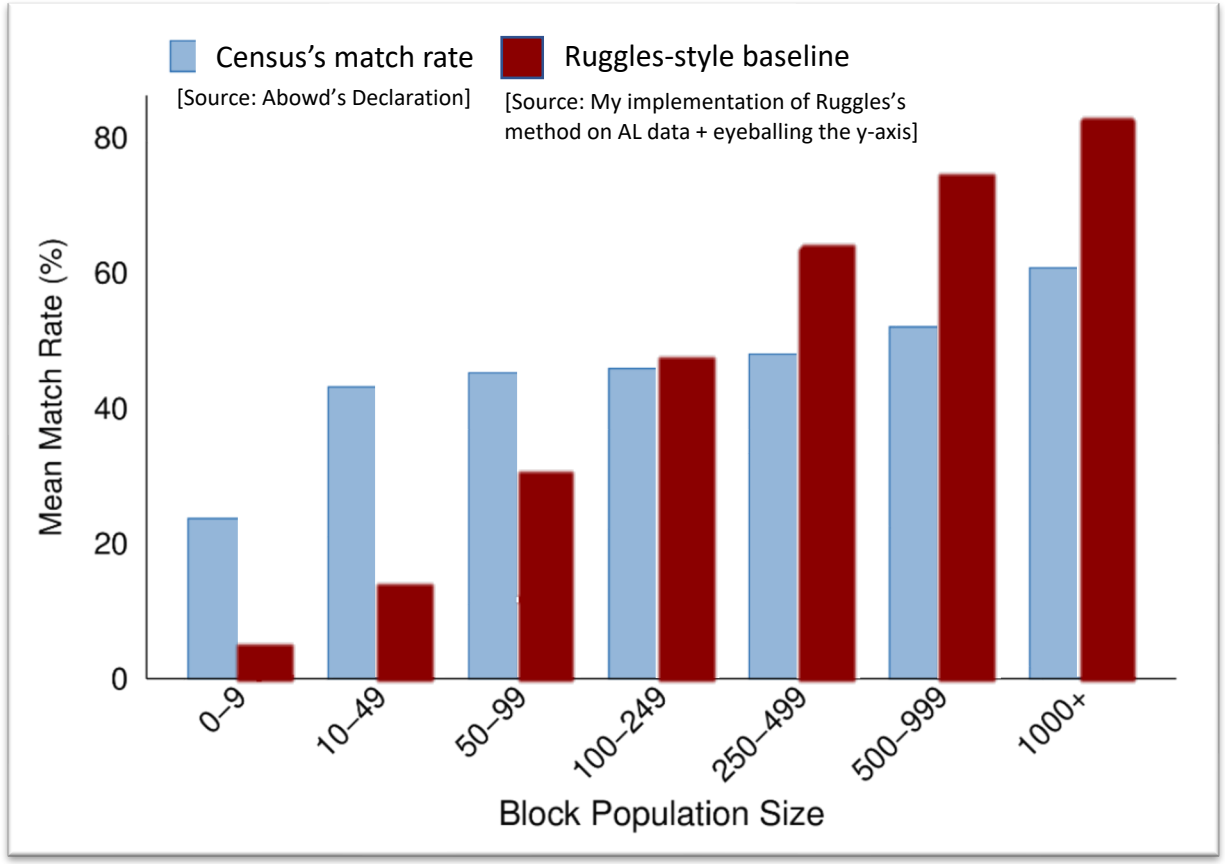
Steven Ruggles @HistDem · Apr 20, 2021

2.I argue that the database reconstruction experiment did not demonstrate a convincing threat to confidentiality, because the results reported by the Census Bureau can be largely explained by chance.

1 1 2

2010 Reconstruction





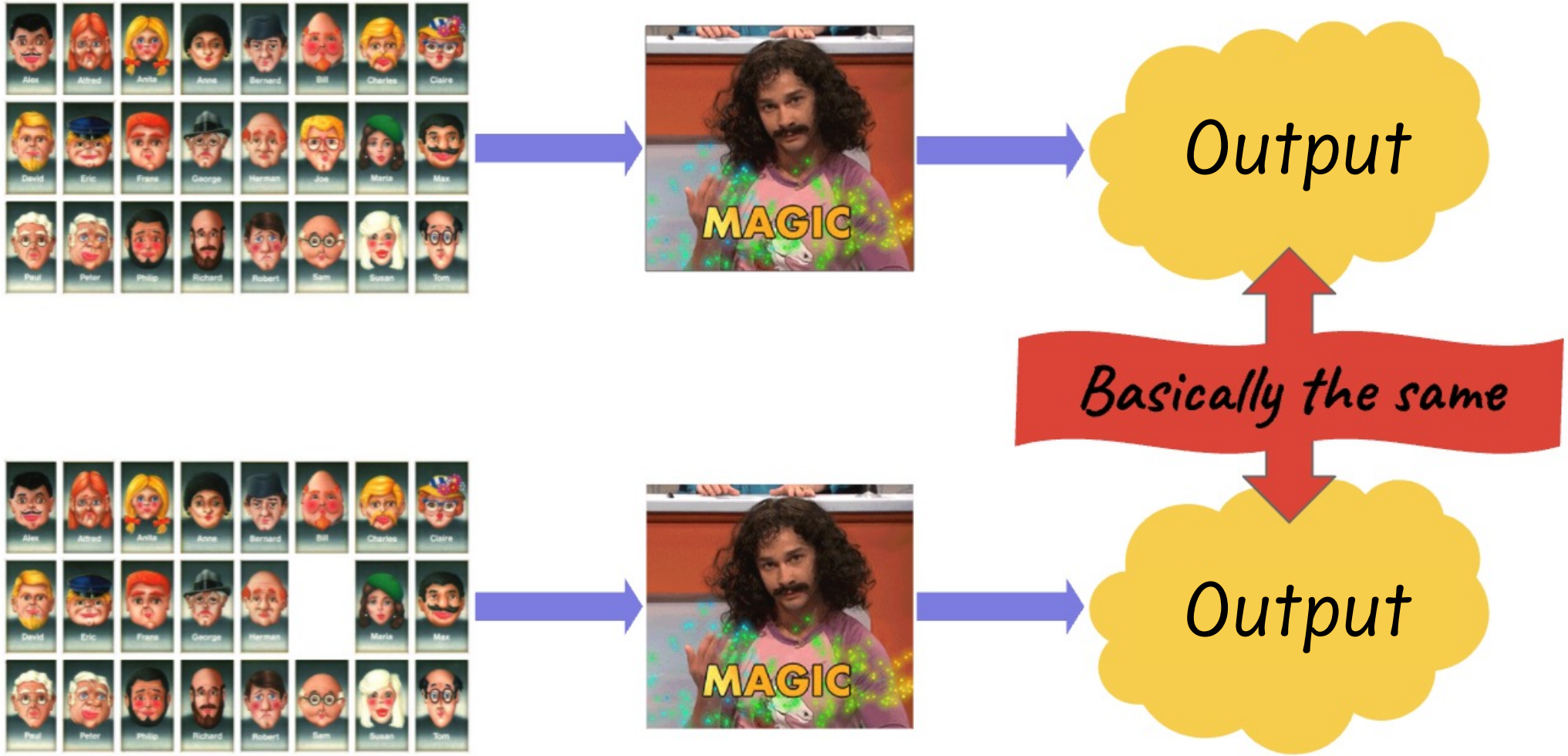
From: Filings in Alabama v Dept of Commerce; Ruggles, Van Riper, "The Role of Chance in the Census Bureau Database Reconstruction Experiment"

2020 Census Operational Plan

A New Design for the 21st Century

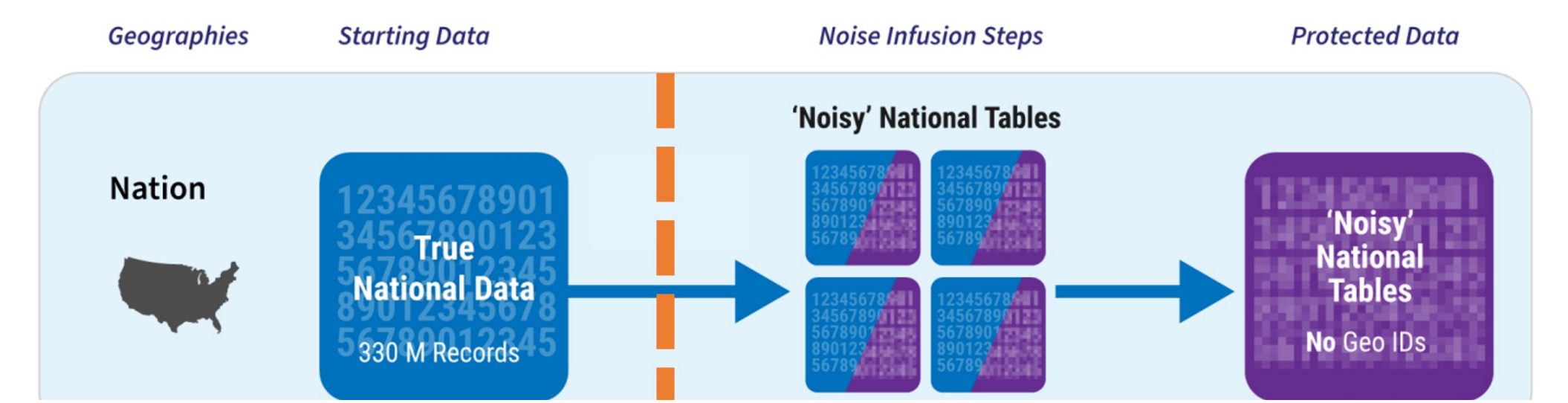
Issued January 2022
Version 5.0

Differential privacy



https://desfontain.es/privacy/differential-privacy-awesomeness.html

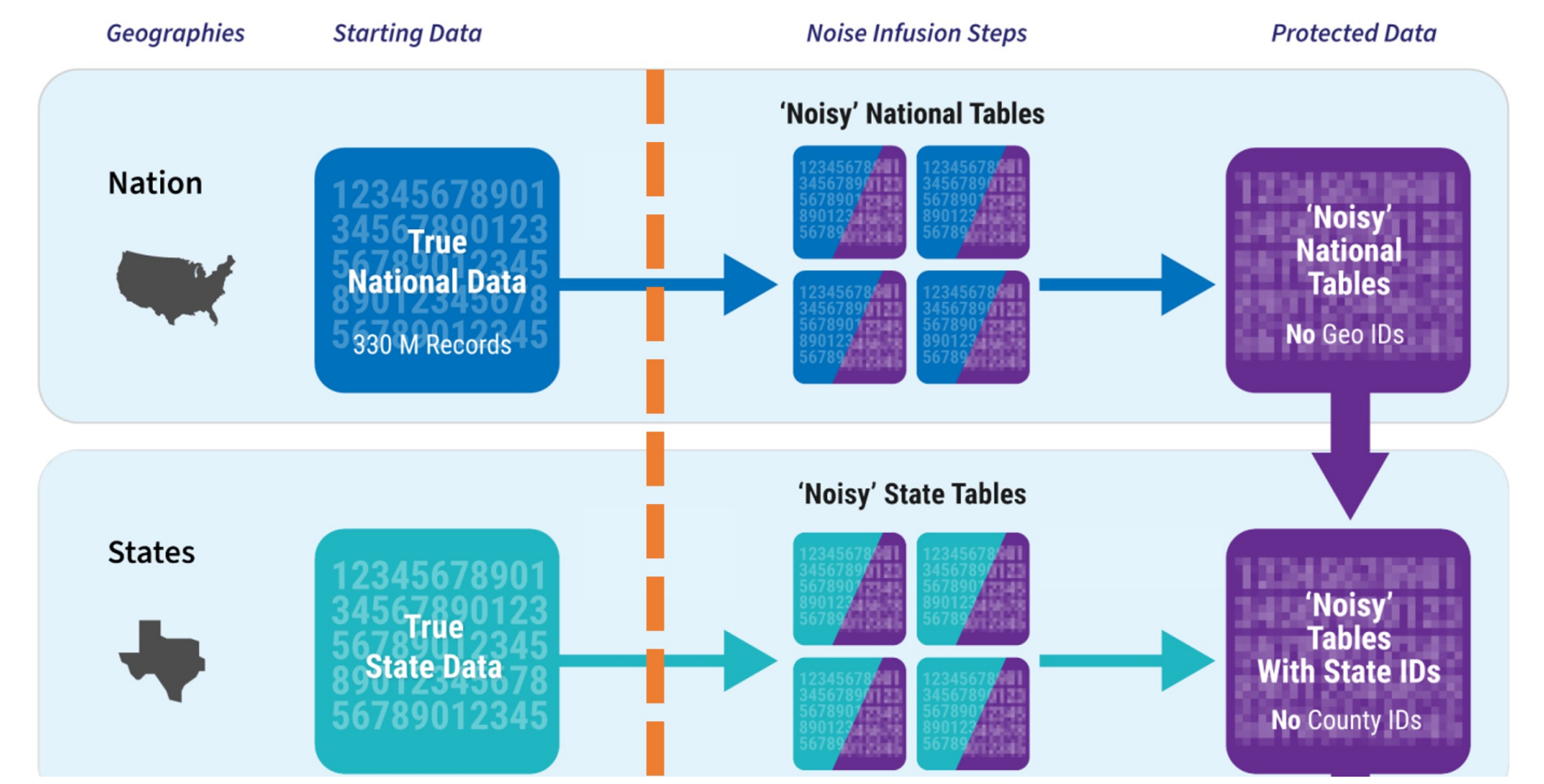
TopDown Algorithm



Differentially private noise infusion

Post processing

TopDown Algorithm



TopDown Algorithm

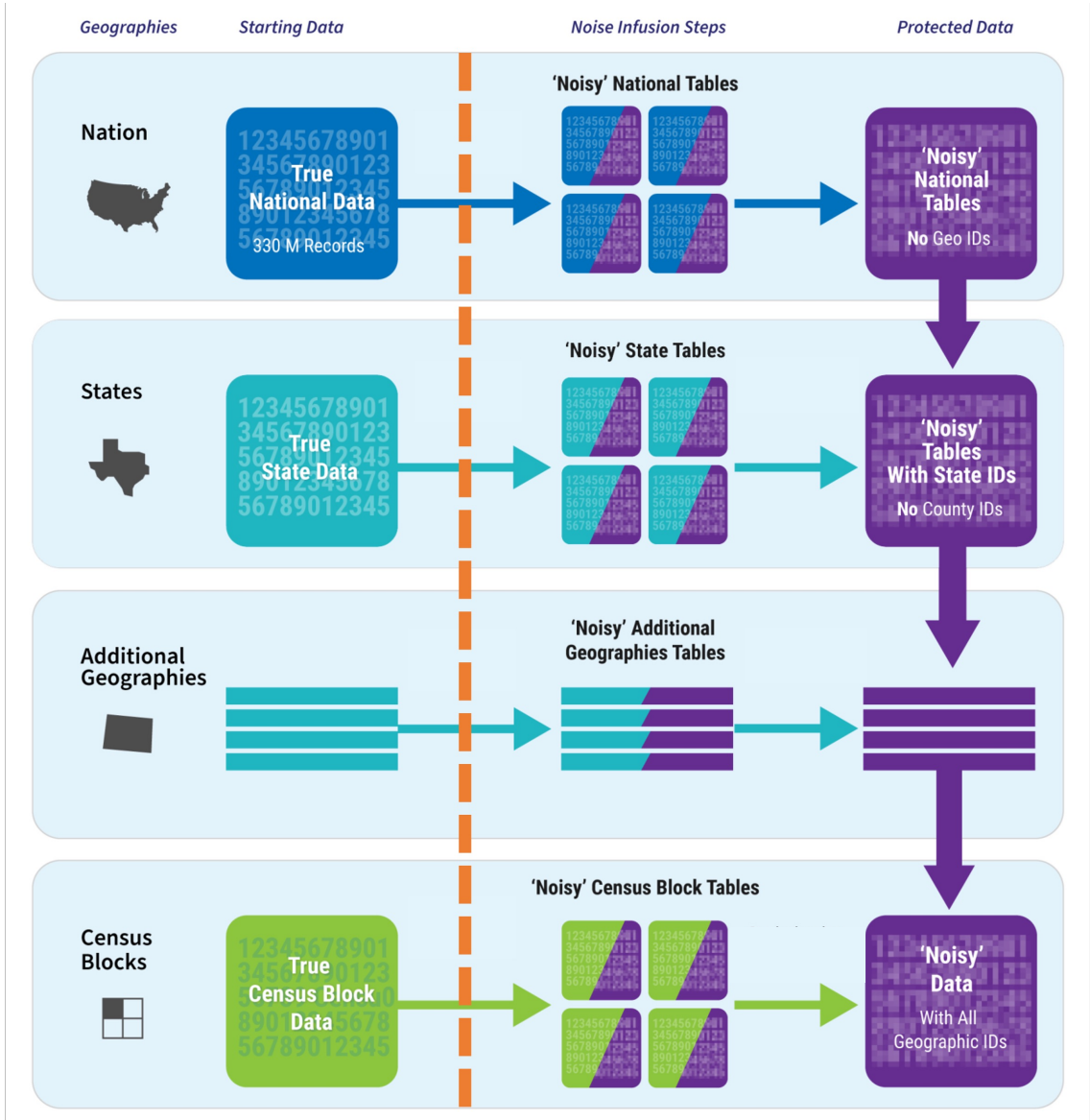


Table 2.3. **Hypothetical Example of Post-Processing**

Step 2: Post-processing

Block	Enumerated counts			Noise			Preliminary noisy counts			Post-processed counts		
	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population	Population under age 18	Population aged 18 and over	Total population
Block 1	25	75	100	0	−4	2	25	71	102	27 (+2)	71 (−4)	98 (−2)
Block 2	20	70	90	−3	2	3	17	72	93	19 (−1)	72 (+2)	91 (+1)
Block 3	10	40	50	2	−3	−2	12	37	48	12 (+2)	37 (−3)	49 (−1)
Block 4	1	9	10	−2	1	1	−1	10	11	0 (−1)	11 (+2)	11 (+1)
Block 5	1	2	3	0	2	0	1	4	3	1 (+0)	4 (+2)	5 (+2)
									Block group	59	195	254

Source: U.S. Census Bureau.

Noise from TopDown is relatively small

Error Statistics for Total Population for Counties (Excluding Puerto Rico)

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**1.75**	**–4**	**+4**
Counties with total population between 0–249	2	2.00	–1	+3
Counties with total population between 250–749	19	1.32	–2	+2
Counties with total population between 750–1,249	26	1.38	–2	+4
Counties with total population between 1,250–1,749	24	1.00	–2	+3
Counties with total population between 1,750–1,949	14	1.14	–1	+2
Counties with total population between 1,950–2,049	10	1.50	–1	+5
Counties with total population between 2,050–2,249	16	0.88	–1	+1
Counties with total population between 2,250–2,749	35	1.31	–2	+3
Counties with total population between 2,750–3,249	38	1.29	–2	+3
Counties with total population at or above 3,250	2,959	1.79	–4	+4

Coverage Error

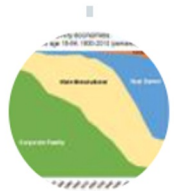
All counties	**3,143**	**964.00**	**–1,841**	**+2,048**

Nonsampling Variability

All counties	**3,143**	**117.27**	**–248**	**+230**

https://www.census.gov/content/dam/Census/library/factsheets/2022/variability.pdf


Apocalypse Now

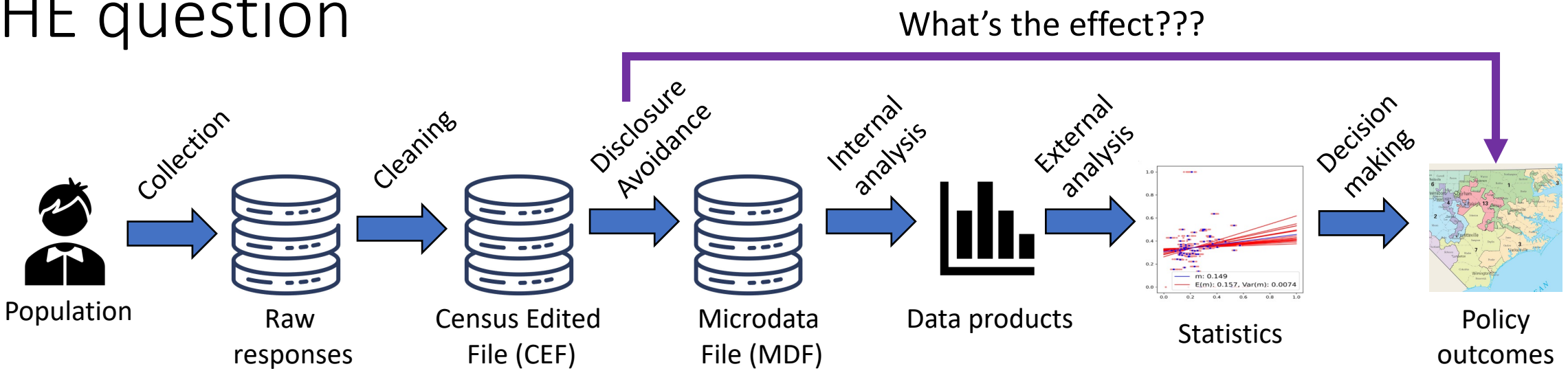


Steven Ruggles @HistDem · Jul 5, 2019

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/

2 12 32

THE question



Redistricting data (PL-94)

Everything Moon just discussed

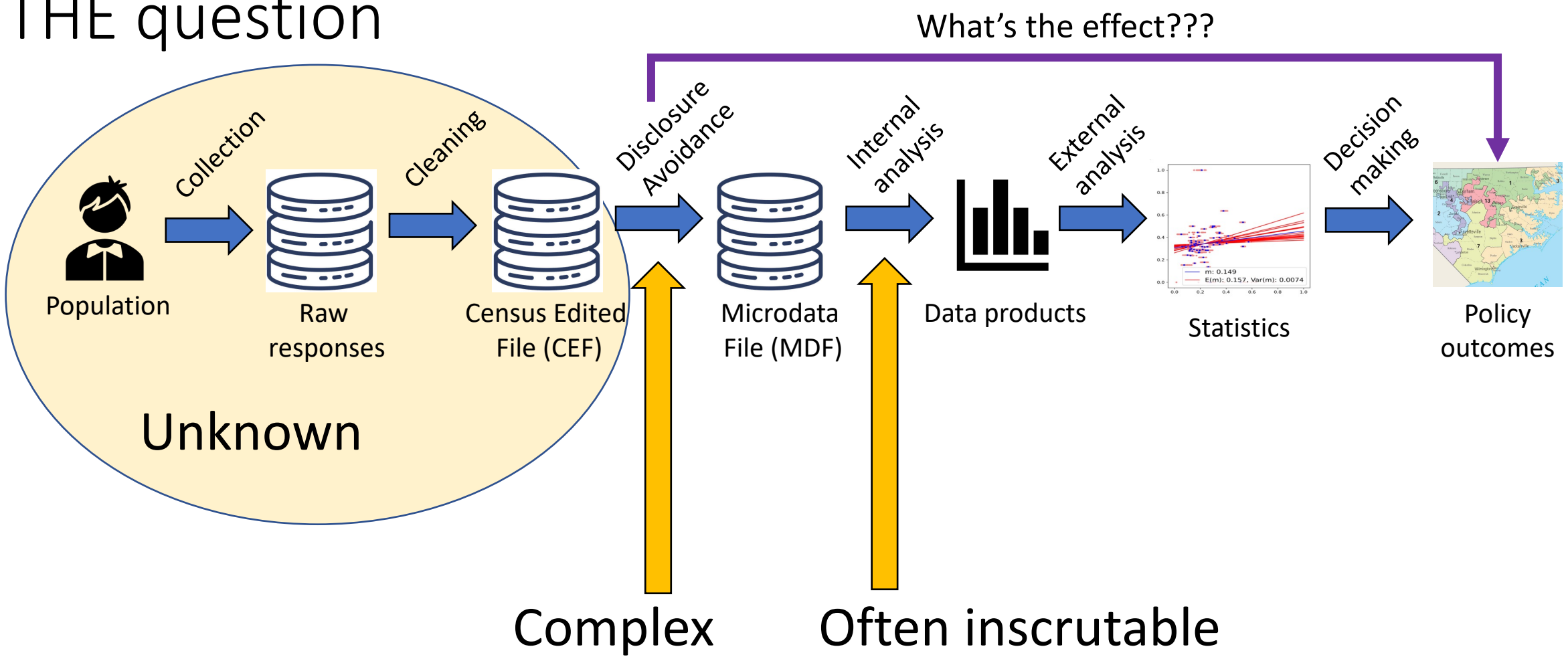
Small Area Income & Poverty Estimates

Title 1 Funding

Sample surveys (eg: ACS)

Social science + policymaking writ large

THE question



The quagmire

Option 0: Be the Census

Table 11V. Counts & Measures of Variation for Tate County School Districts, MS Twenty-five Runs of the *TDA* for County Districts 01, 02, 03, 04, 05 ($C_T(g)$ counts result from 2020 Census Redistricting Data Production Settings ($\epsilon = 17.14$ for persons) version of *TDA*.)

	(Measures of Variation)											
DIST-ID	Tate Schools		01	01	02	02	03	03	04	04	05	05
	$\bar{C}_T(g)$	$C_S(g)$	$\bar{C}_T(g)$	$C_S(g)$	$\bar{C}_T(g)$	$C_S(g)$	$\bar{C}_T(g)$	$C_S(g)$	$\bar{C}_T(g)$	$C_S(g)$	$\bar{C}_T(g)$	$C_S(g)$
	$\sqrt{V(1)_g}$	$\sqrt{V(2)_g}$	$\sqrt{V(1)_g}$	$\sqrt{V(2)_g}$	$\sqrt{V(1)_g}$	$\sqrt{V(2)_g}$	$\sqrt{V(1)_g}$	$\sqrt{V(2)_g}$	$\sqrt{V(1)_g}$	$\sqrt{V(2)_g}$	$\sqrt{V(1)_g}$	$\sqrt{V(2)_g}$
Demographic (g)	$RV(1)_g$	$RV(2)_g$	$RV(1)_g$	$RV(2)_g$	$RV(1)_g$	$RV(2)_g$	$RV(1)_g$	$RV(2)_g$	$RV(1)_g$	$RV(2)_g$	$RV(1)_g$	$RV(2)_g$
TOTAL	18,815	18,823	3,916	3,914	3,885	3,893	3,644	3,665	3,714	3,697	3,657	3,654
	18	20	22	22	21	23	20	30	26	31	16	16
	0.001	0.001	0.006	0.006	0.005	0.006	0.006	0.008	0.007	0.008	0.004	0.004
TOTAL18	13,892	13,893	2,776	2,780	2,833	2,826	2,789	2,799	2,766	2,755	2,728	2,733
	17	17	20	21	19	20	14	17	23	26	13	14
	0.001	0.001	0.007	0.007	0.007	0.007	0.005	0.006	0.008	0.009	0.005	0.005
TOTALHISP	423	399	95	87	64	63	106	110	51	32	106	107
	9	26	6	10	4	4	8	9	6	20	8	8
	0.021	0.064	0.066	0.118	0.063	0.066	0.073	0.078	0.119	0.631	0.072	0.071
TOTALNH	18,392	18,424	3,821	3,827	3,821	3,830	3,537	3,555	3,663	3,665	3,551	3,547
	18	37	22	23	21	23	19	26	24	24	18	18
	0.001	0.002	0.006	0.006	0.005	0.006	0.005	0.007	0.007	0.007	0.005	0.005
WHITENH	12,805	12,841	3,387	3,378	1,613	1,628	2,833	2,860	2,276	2,293	2,696	2,682
	13	39	14	17	14	21	14	30	20	26	12	19
	0.001	0.003	0.004	0.005	0.009	0.013	0.005	0.011	0.009	0.011	0.005	0.007
BLACKNH	5,394	5,389	373	400	2,158	2,139	678	666	1,363	1,349	822	835
	11	12	12	30	10	21	11	16	14	20	15	20
	0.002	0.002	0.033	0.074	0.004	0.010	0.016	0.024	0.010	0.015	0.018	0.024

From: Wright, Irimata, "Empirical Study of Two Aspects of the Topdown Algorithm Output for Redistricting: Reliability & Variability (August update)"

Option 1: A wrong, but useful model

- Noise $\mathrm{Z} \sim N\left(\frac{1}{\rho^*}\right)$ added to counts in each county / tract / block
- Non-negativity + consistency
  → Small counts biased upwards
  → Large counts biased downwards

$$\mathbb{E}(|Z_h|) \approx \frac{0.8}{\sqrt{\rho}}$$

$\rho_{pop,county} = 0.213 \rightarrow 1.73$

$\rho_{pop,tract} = 0.164 \rightarrow 1.97$

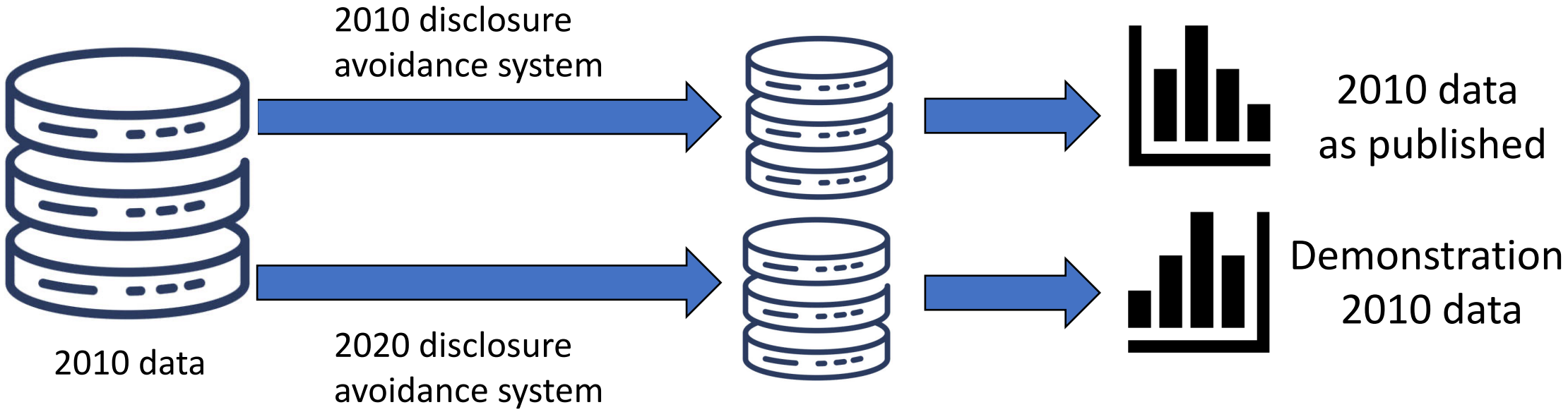
Privacy-Loss Budget: People

Geographic level	Rho allocation
United States	104/4,099
State	1,440/4,099
County	447/4,099
Tract	687/4,099
Optimized block group[1]	1,256/4,099
Block	165/4,099

B	C	D	E	F	G
name	state	TOTPOP_dp	TOTPOP_sf	|Deviation|	Mean(|Deviation|)
Autauga County	1	54574	54571	3	1.741695126
Baldwin County	1	182266	182265	1	
Barbour County	1	27456	27457	1	
Bibb County	1	22917	22915	2	
Blount County	1	57322	57322	0	
Bullock County	1	10915	10914	1	

B	C	D	E	F	G
name	state	TOTPOP_dp	TOTPOP_sf	|Deviation|	Mean(|Deviation|)
Census Tract 201	1	1910	1912	2	1.924650685
Census Tract 202	1	2171	2170	1	
Census Tract 203	1	3371	3373	2	
Census Tract 204	1	4384	4386	2	

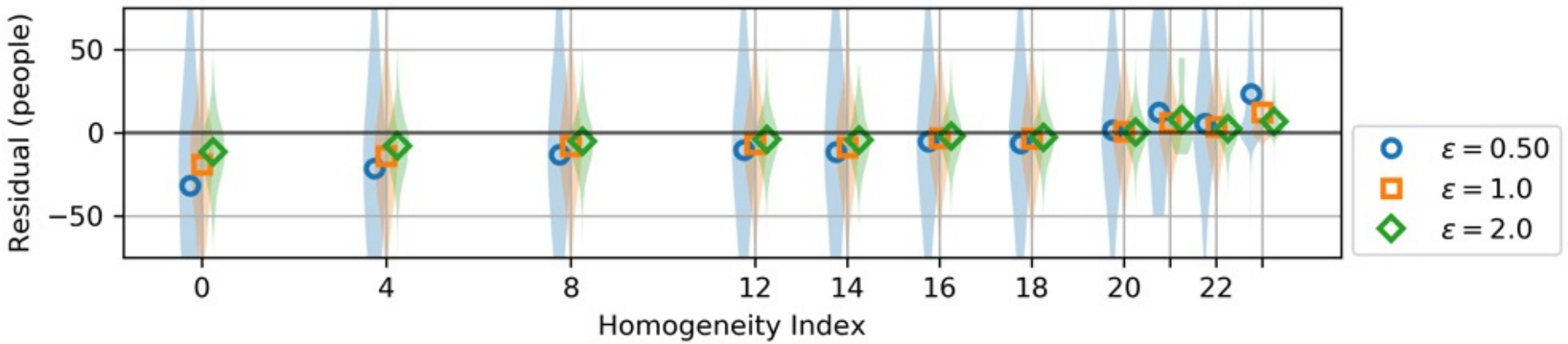
Option 2: Demonstration data



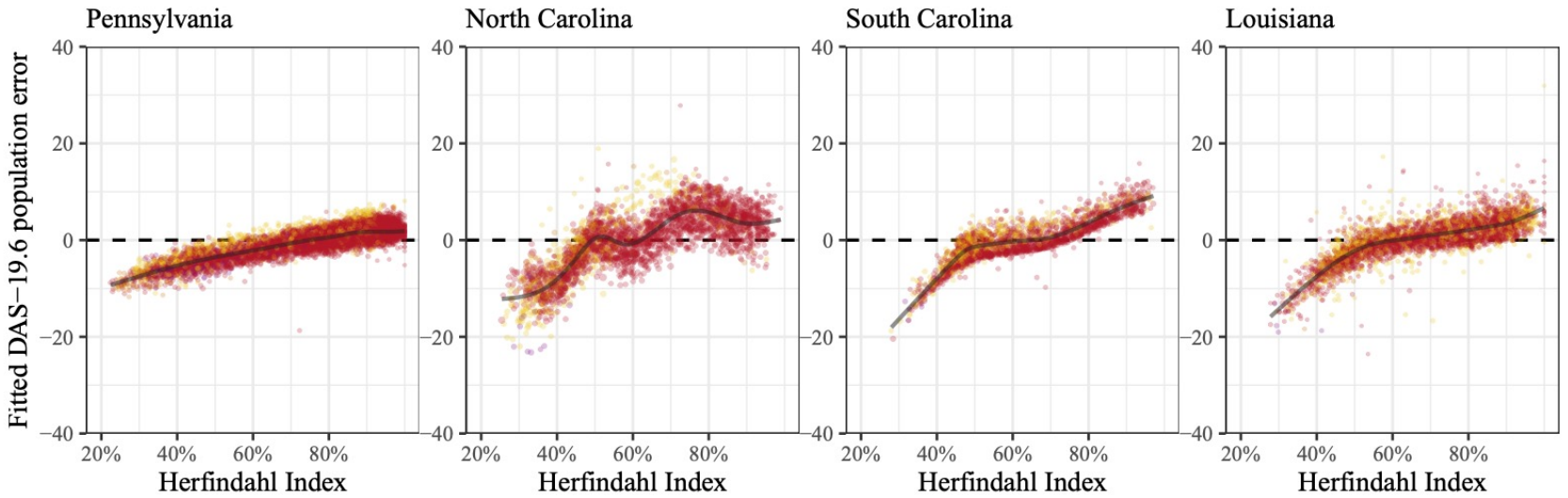
- Descriptive statistics
  - Learn about (TopDown – Swapping)
- Redo prior analysis and compare
  - Learn about effect of noise generally

Bias from non-negativity

1940 data
old TopDown

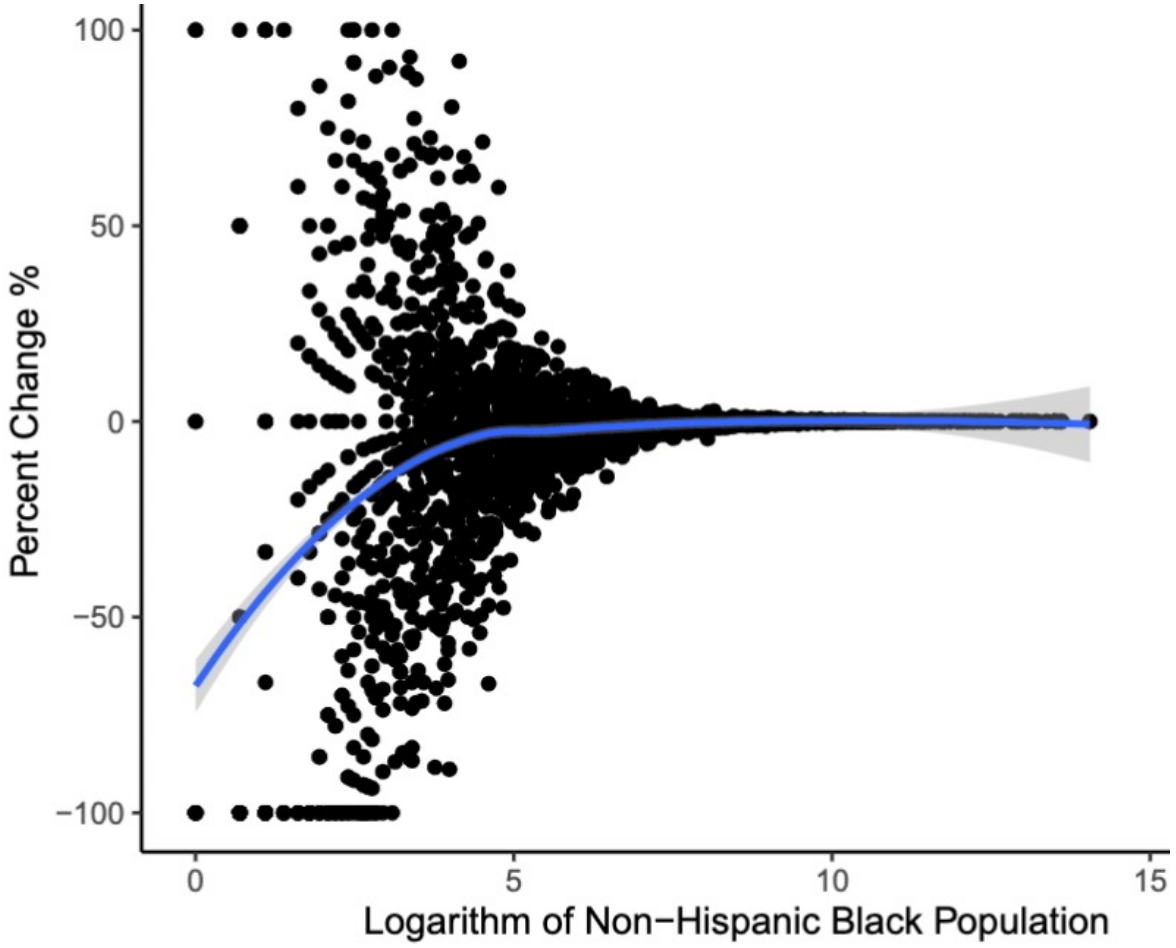


2010
production
TopDown



From: Petti, Flaxman “Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff”
Kenny, Kuriwaki, McCartan, Rosenman, Simko, Imai.``The Use of Differential Privacy for Census Data and its Impact on Redistricting: The Case of the 2020 U.S. Census.''

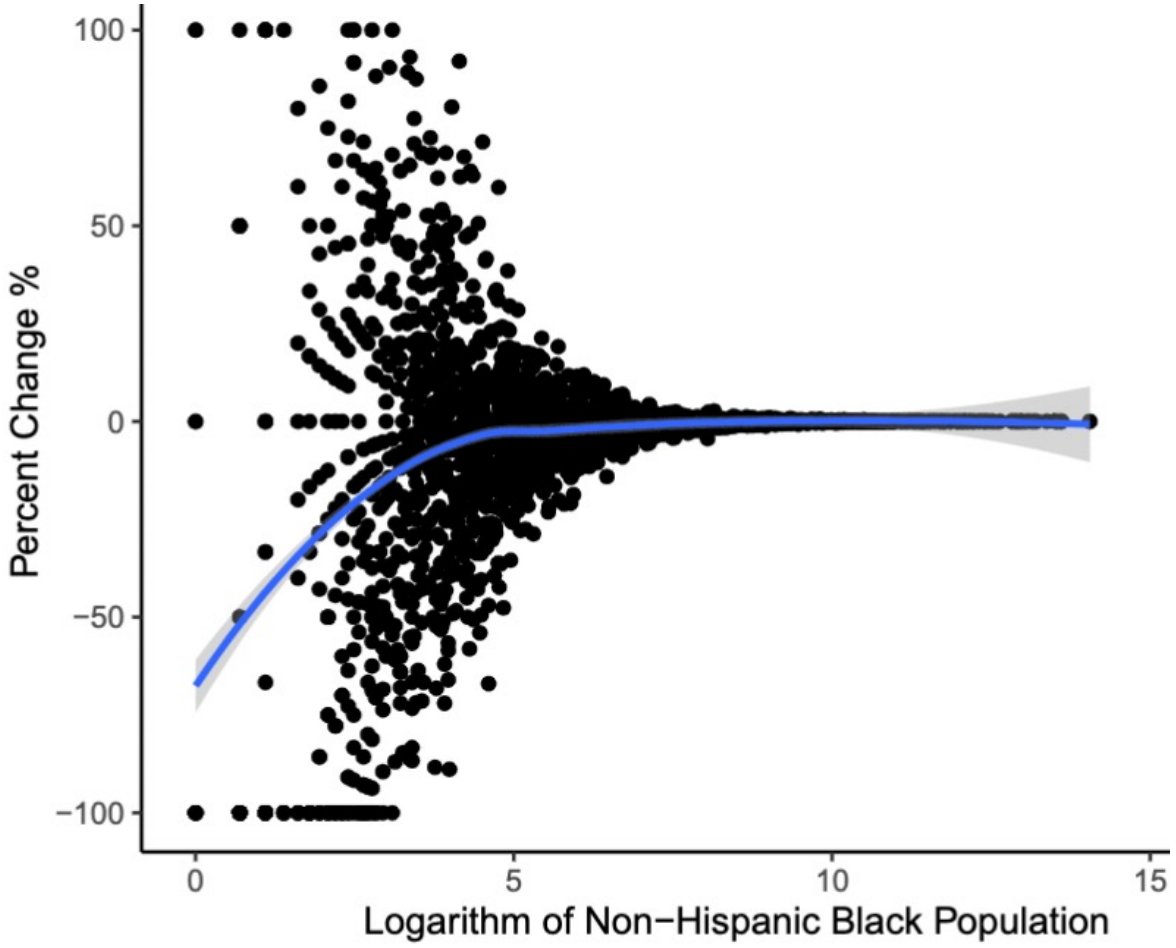
Mortality rates



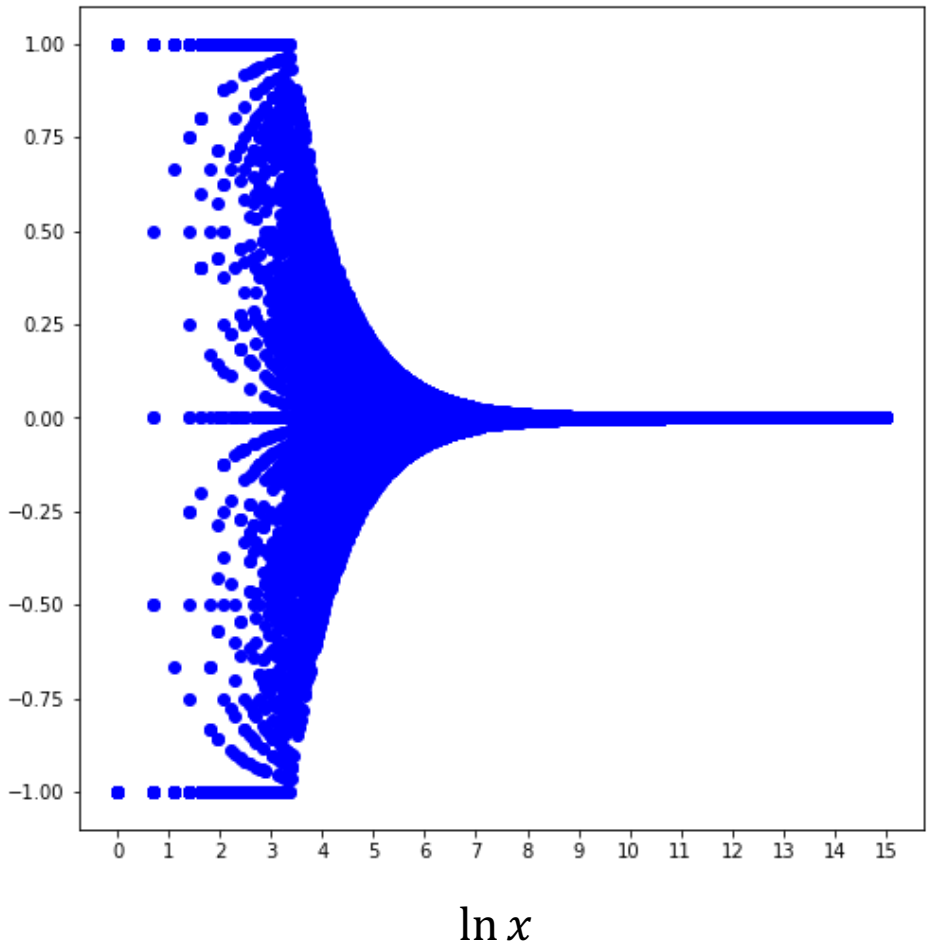
We find that the implementation of differential privacy will produce **dramatic changes in population counts for racial/ethnic minorities in small areas** and less urban settings, significantly altering knowledge about health disparities in mortality.

Santos-Lozada et al, https://www.pnas.org/doi/epdf/10.1073/pnas.2003714117

Mortality rates

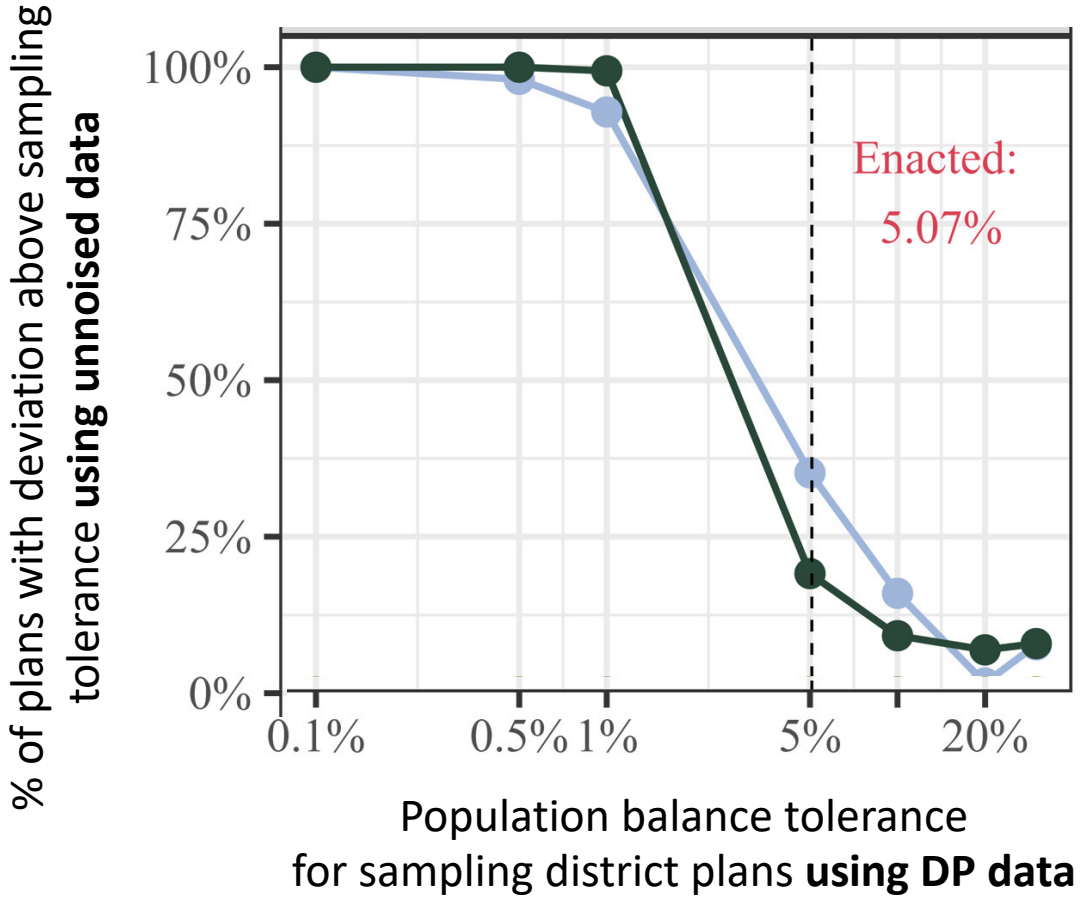


$$\frac{Unif(\pm 30)}{x}$$



Santos-Lozada et al, https://www.pnas.org/doi/epdf/10.1073/pnas.2003714117

Population-balanced redistricting



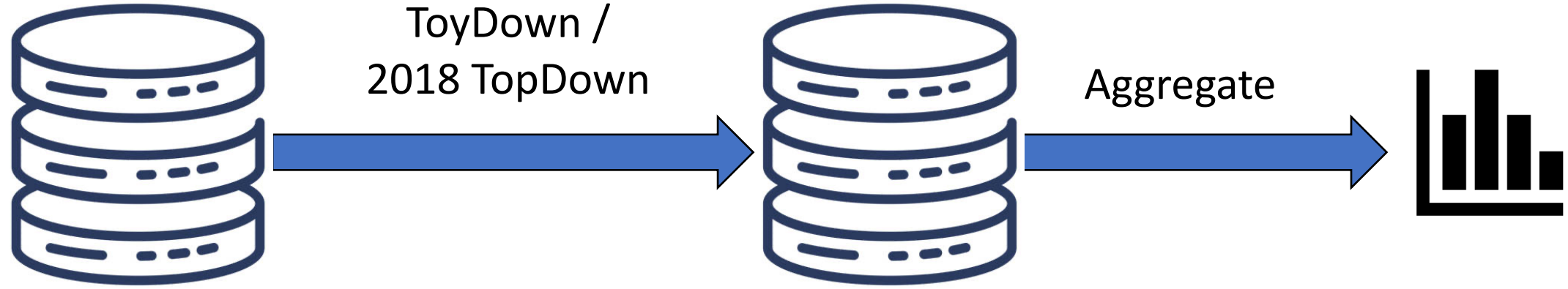
"We conduct our empirical evaluation under a likely scenario, in which practitioners, map drawers and **analysts** alike, **treat these DAS-protected data 'as is'** as they have done in the past, **without accounting for the DAS noise** generation mechanism.

"Our analysis shows that the added noise makes it **impossible to follow the principle of One Person, One Vote, "**

Solution: set sampling tolerance ≈0.3% lower than policy tolerance

Kenny et al, https://www.science.org/doi/10.1126/sciadv.abk3283

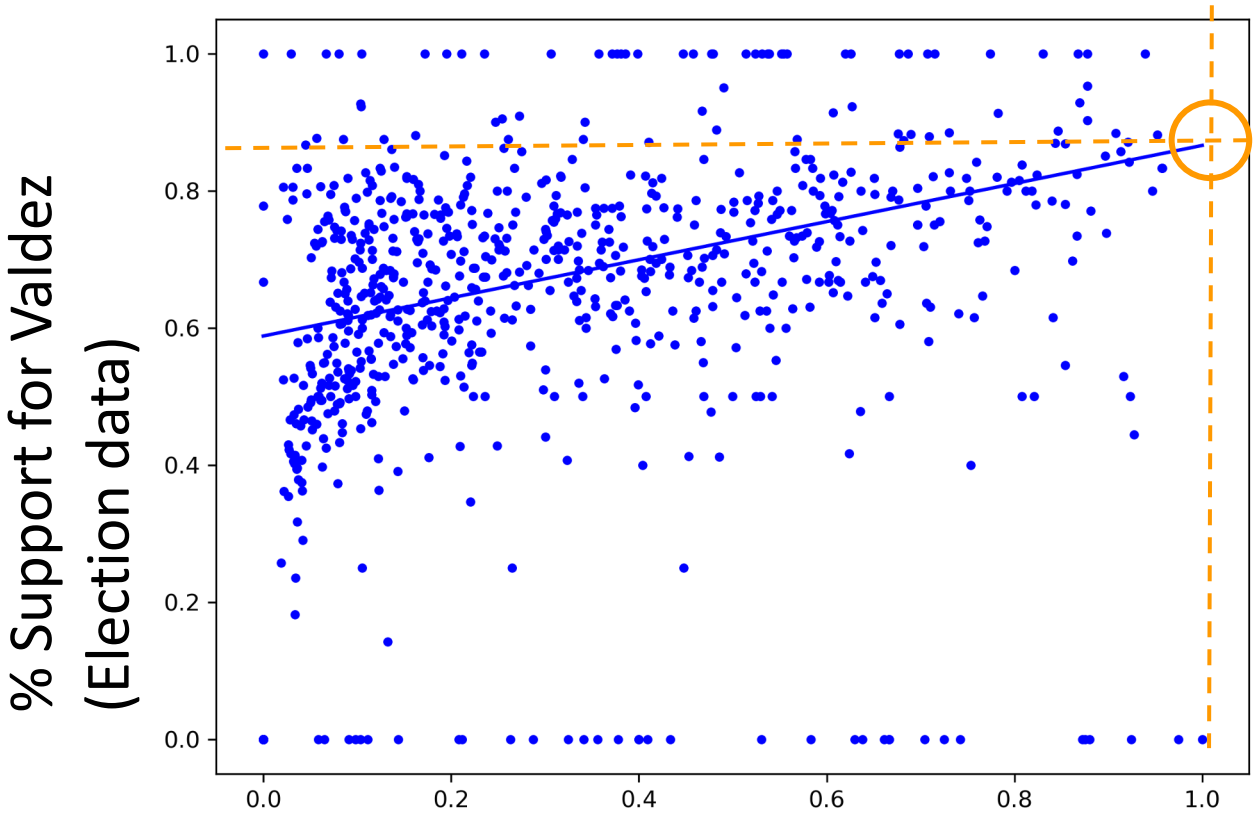
Option 3: Try to run TopDown



From: Cohen, Duchin, Matthews, Suwal, “Private Numbers in Public Policy: Census, Differential Privacy, and Redistricting“

Measuring racial polarized voting (RPV)

Standard ecological regression

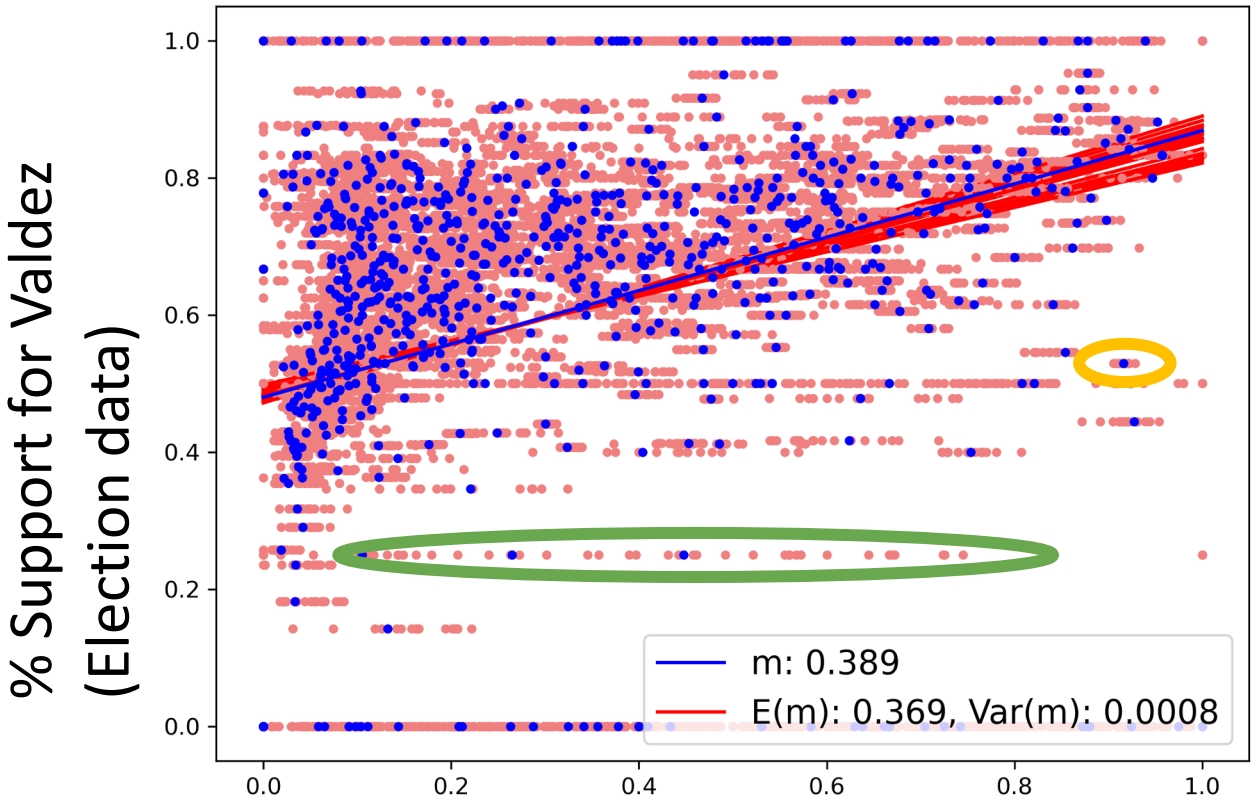


Support for Lupe Valdez in the 2018 Governor election in Texas across the 827 precincts in Dallas County.

Each dot is a precinct, and the blue line is the line of best fit.

RPV with noise

Standard ecological regression



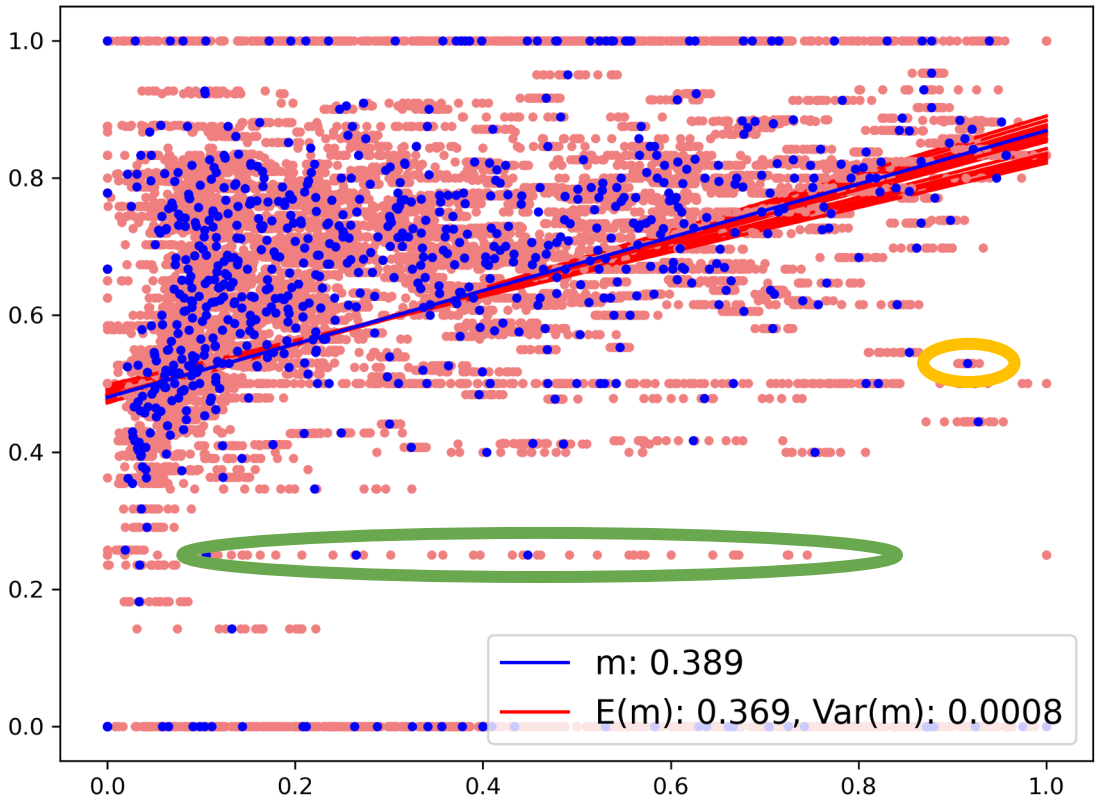
Support for Lupe Valdez in the 2018 Governor election in Texas across the 827 precincts in Dallas County.

Each dot is a precinct, and the blue line is the line of best fit.

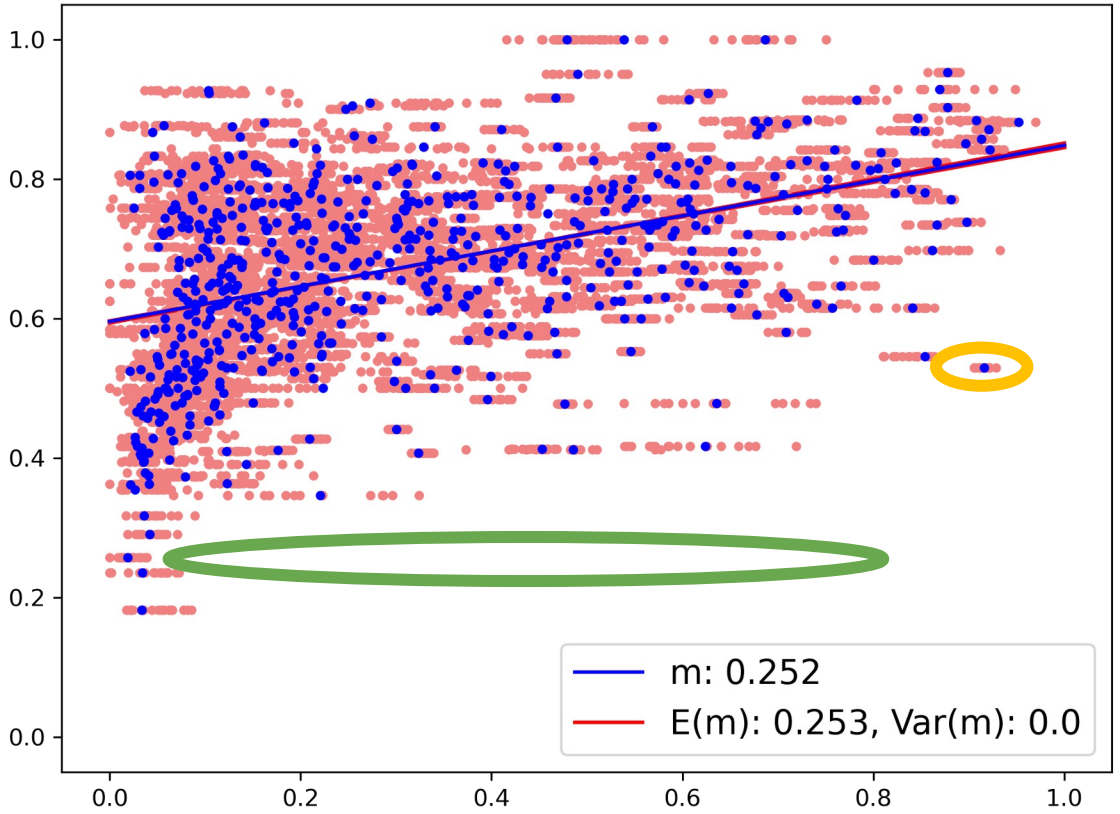
Pink = Noised Numbers

Accounting for small precincts

Standard ecological regression



Filtering precincts with < 10 votes
(24% of precincts!)

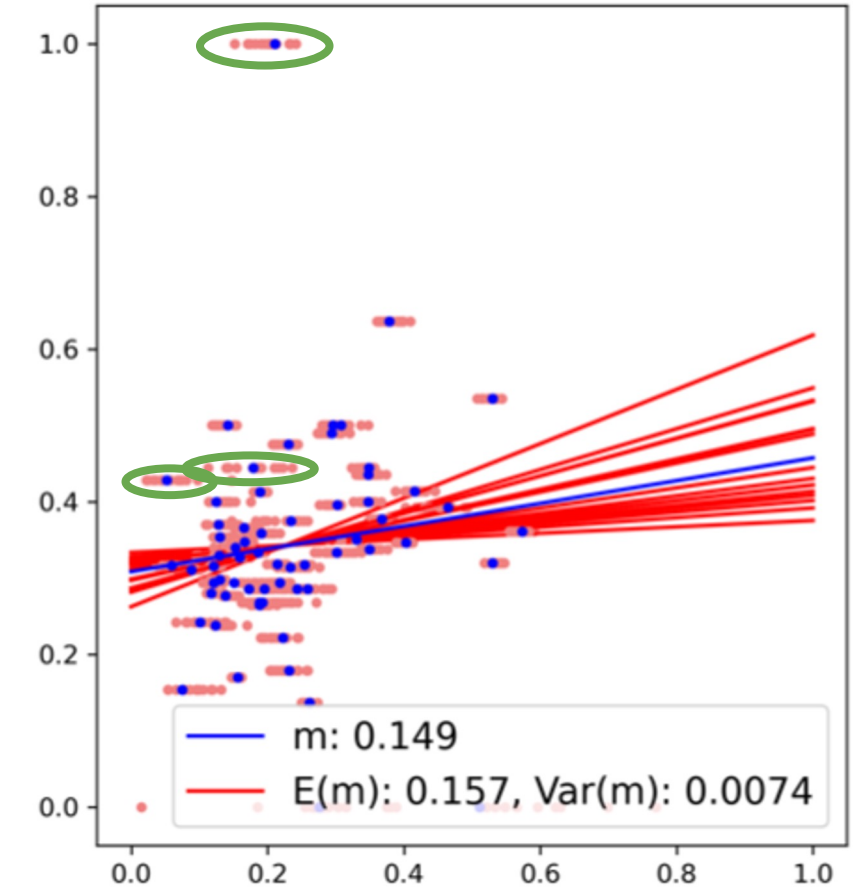


% Support for Valdez
(Election data)

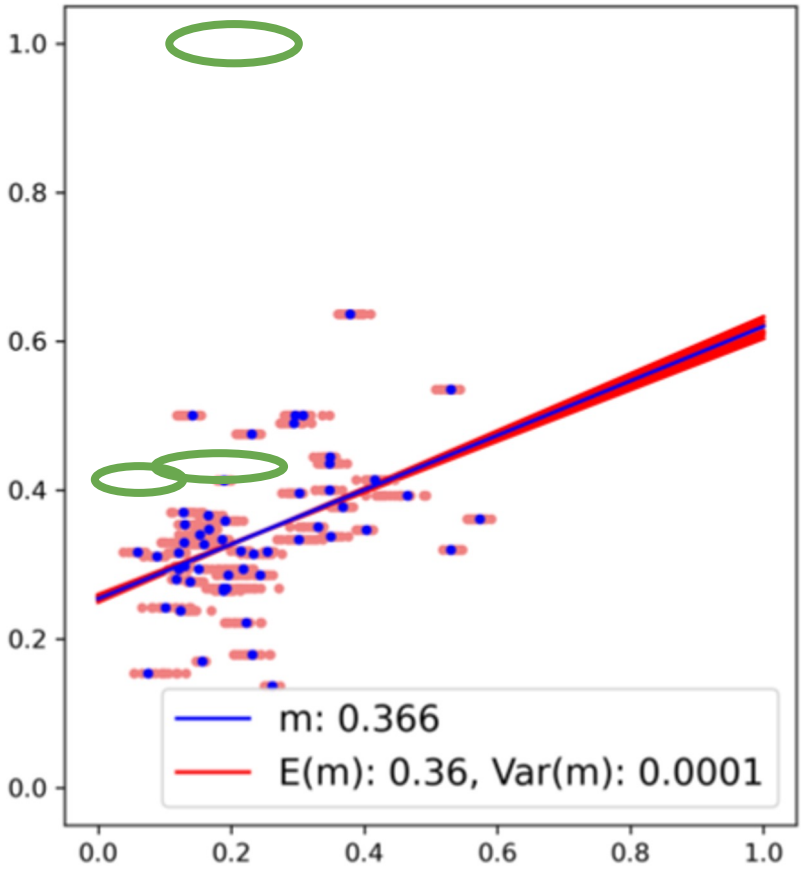
% Hispanic Population
(Census data)

Accounting for small precincts

Standard ecological regression



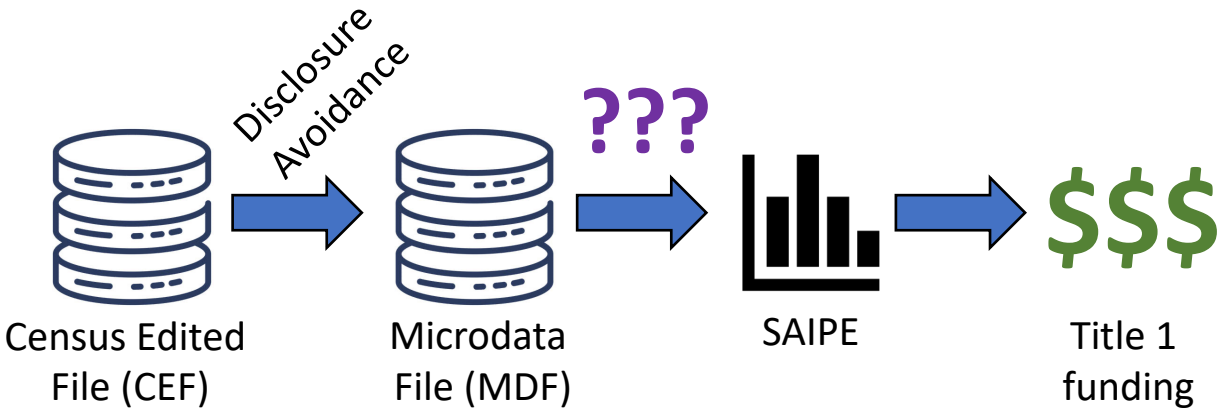
Filtering precincts with < 10 votes

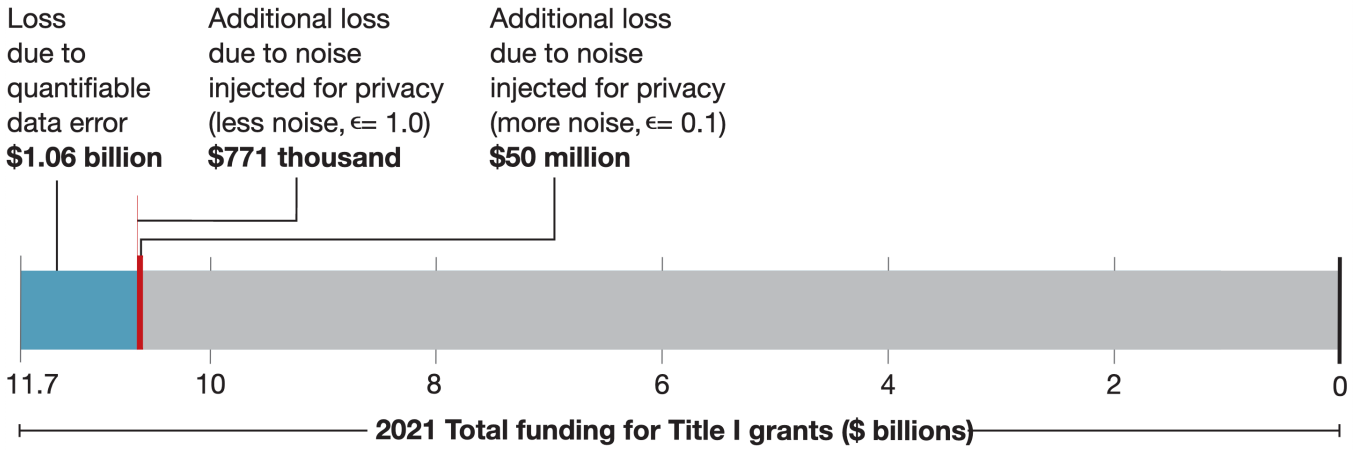


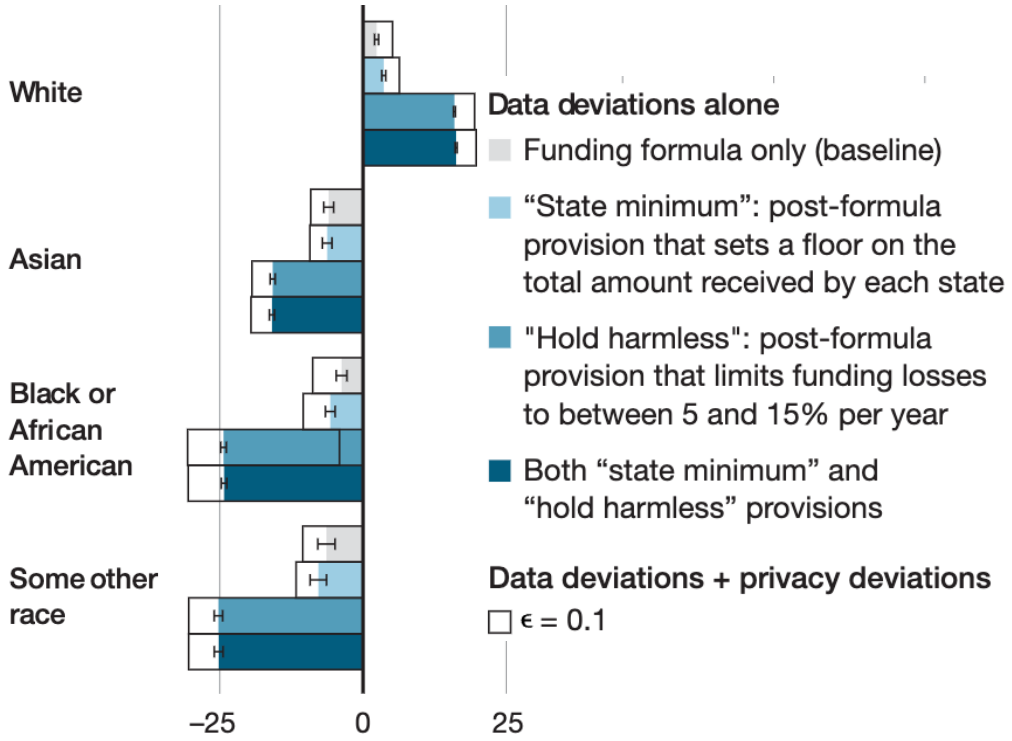
% Support for Valdez
(Election data)

% Hispanic Population
(Census data)

Option 4: Go straight to the policy question







Race-weighted misallocation per eligible child ($)
[edited, partial figure]

From: Steed, Liu, Wu, Acquisti "Policy impacts of statistical uncertainty and privacy"

UNITED STATES DISTRICT COURT FOR THE MIDDLE DISTRICT OF ALABAMA EASTERN DIVISION

THE STATE OF ALABAMA; ROBERT ADERHOLT, Representative for Alabama's 4th Congressional District, in his official and individual capacities; WILLIAM GREEN; and CAMARAN WILLIAMS,

Plaintiffs,

v.

UNITED STATES DEPARTMENT OF COMMERCE; GINA RAIMONDO, in her official capacity as Secretary of Commerce; UNITED STATES BUREAU OF THE CENSUS, an agency within the United States Department of Commerce; and RON JARMIN, in his official capacity as Acting Director of the U.S. Census Bureau,

Defendants.

CIVIL ACTION NO. _____

COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

THREE-JUDGE COURT REQUESTED PURSUANT TO 28 U.S.C. § 2284

INTRODUCTION

1. This suit challenges two unlawful actions by the U.S. Commerce Department and Census Bureau in relation to the 2020 decennial census—(1) Defendants' decision to produce manipulated redistricting data to the States, and (2) Defendants' refusal to produce redistricting data on time.

2. First, the skewed numbers. Congress has ordered the Secretary of Commerce to work with the States to learn what they need for redistricting and then report to each State accurate "[t]abulations of population" for subparts of each State for purposes of "legislative apportionment or districting of such State." 13 U.S.C. § 141(c). But the Secretary, through the Census Bureau, has announced that she will instead provide the States purposefully flawed population tabulations. The Bureau intends to use a novel statistical method called differential privacy to intentionally skew the population tabulations provided to States to use for redistricting. Thus, while the Bureau touts its mission "to count everyone once, only once, and in the right place,"[1] it will force Alabama to redistrict using results that purposefully count people in the wrong place.

BRIEF OF *AMICI CURIAE* STATE OF UTAH AND 15 OTHER STATES IN SUPPORT OF PLAINTIFFS

The States of Utah, Alaska, Arkansas, Florida, Kentucky, Louisiana, Maine, Mississippi, Montana, Nebraska, New Mexico, Ohio, Oklahoma, South Carolina, Texas, and West Virginia (*Amici* States) agree with Plaintiffs that the Secretary's intended use of differential privacy deprives states of accurate "[t]abulations of population" of state subparts to use in legislative apportionment and districting under 13 U.S.C. § 141(c). *Amici* States also agree that the Secretary can

Legal question 1:
Was swapping private enough?

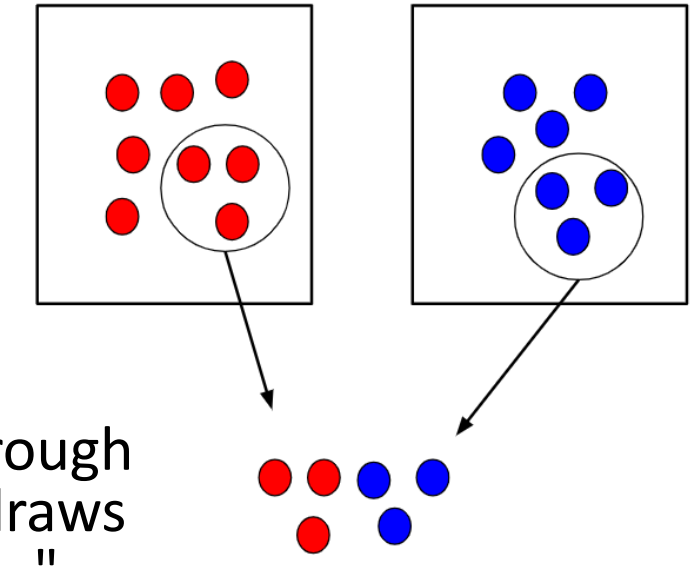
- **Law**
  - The Census Bureau may not “disclose the information reported by, or on behalf of, any particular respondent”
- **Alabama**
  - **“**there is not a single documented case of anyone outside the Census Bureau revealing the responses of a particular identified person in public use decennial census”
- **Census Bureau**
  - “swapping and top and bottom coding applied at the level used in the 2010 census are insufficient to prevent re-identification given the ability to perform database reconstruction and the availability of external data."

Legal question 2:
Does TopDown violate “one person, one vote”?

- **Law**
  - Congressional districts ”as nearly of equal population as is practicable”
    - *Reynolds v Sims (1964)*
- **Alabama**
  - "Congressional districts drawn from the demonstration data would likely violate one-person, one-vote"
- **Census**
  - "[T]he ‘good-faith effort to achieve population equality’ required of a State conducting intrastate redistricting does not translate into a requirement that the Federal Government conduct a census that is as accurate as possible." - *Wisconsin v New York (1996)*

Legal question 3:
Is TopDown an illegal “statistical method”?

- **Law**
  - "the term ‘statistical method’ means ... [any] statistical procedure ... to add or subtract counts to or from the enumeration of the population as a result of statistical inference"
- **Alabama**
  - "Privacy is introduced ... by introducing random error through sampling from statistical distributions.... These random draws are then added or subtracted to the actual observations..."
- **Census**
  - DP not "a result of statistical inference", and "thus unlike the sampling" at issue in precedent



IN THE UNITED STATES DISTRICT COURT
FOR THE MIDDLE DISTRICT OF ALABAMA
EASTERN DIVISION

THE STATE OF ALABAMA, *et al.*,	)	
	)	
Plaintiffs,	)	
	)	
v.	)	Case No. 3:21-cv-211-RAH-ECM-KCN
	)	(WO)
UNITED STATES DEPARTMENT	)	
OF COMMERCE, *et al.*,	)	
	)	
Defendants.	)	

MEMORANDUM OPINION AND ORDER

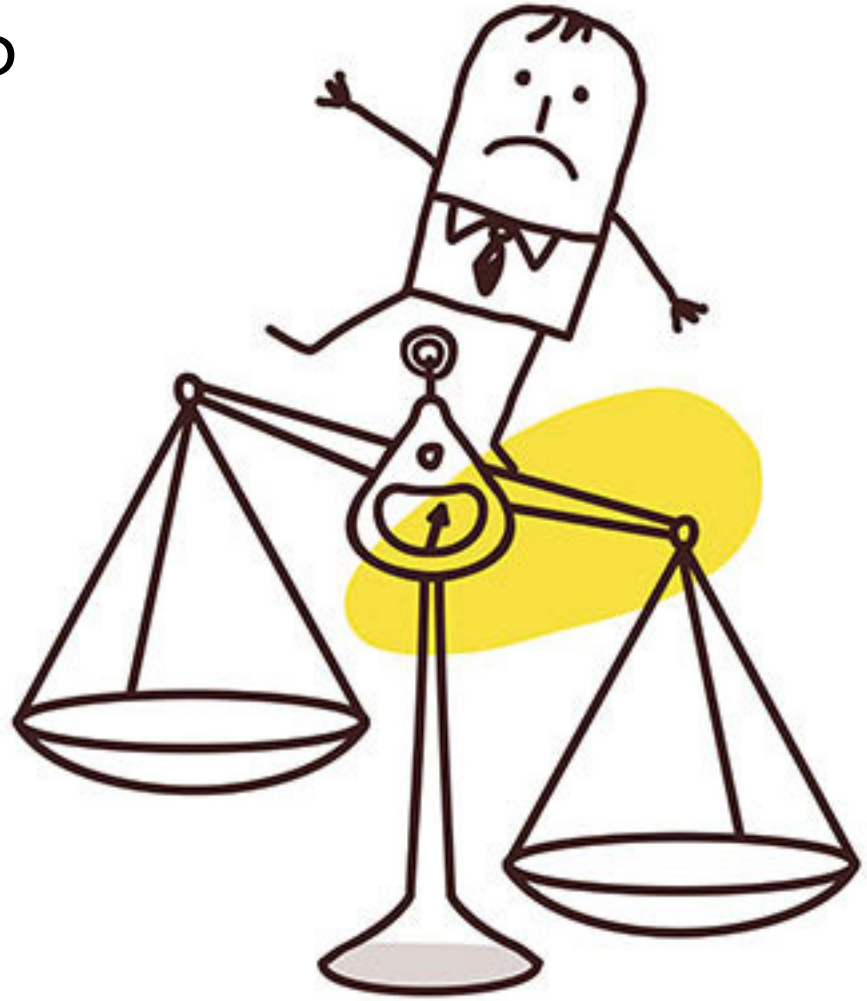
After the benefit of oral argument, the court concludes that Plaintiffs' motion for a preliminary injunction and petition for writ of mandamus are due to be DENIED.

Translation:
Census wins

The outcome had little to do with DP

Alabama needs **standing** to sue

- Injury in fact
- Traceable
- Redress
- Right of Action

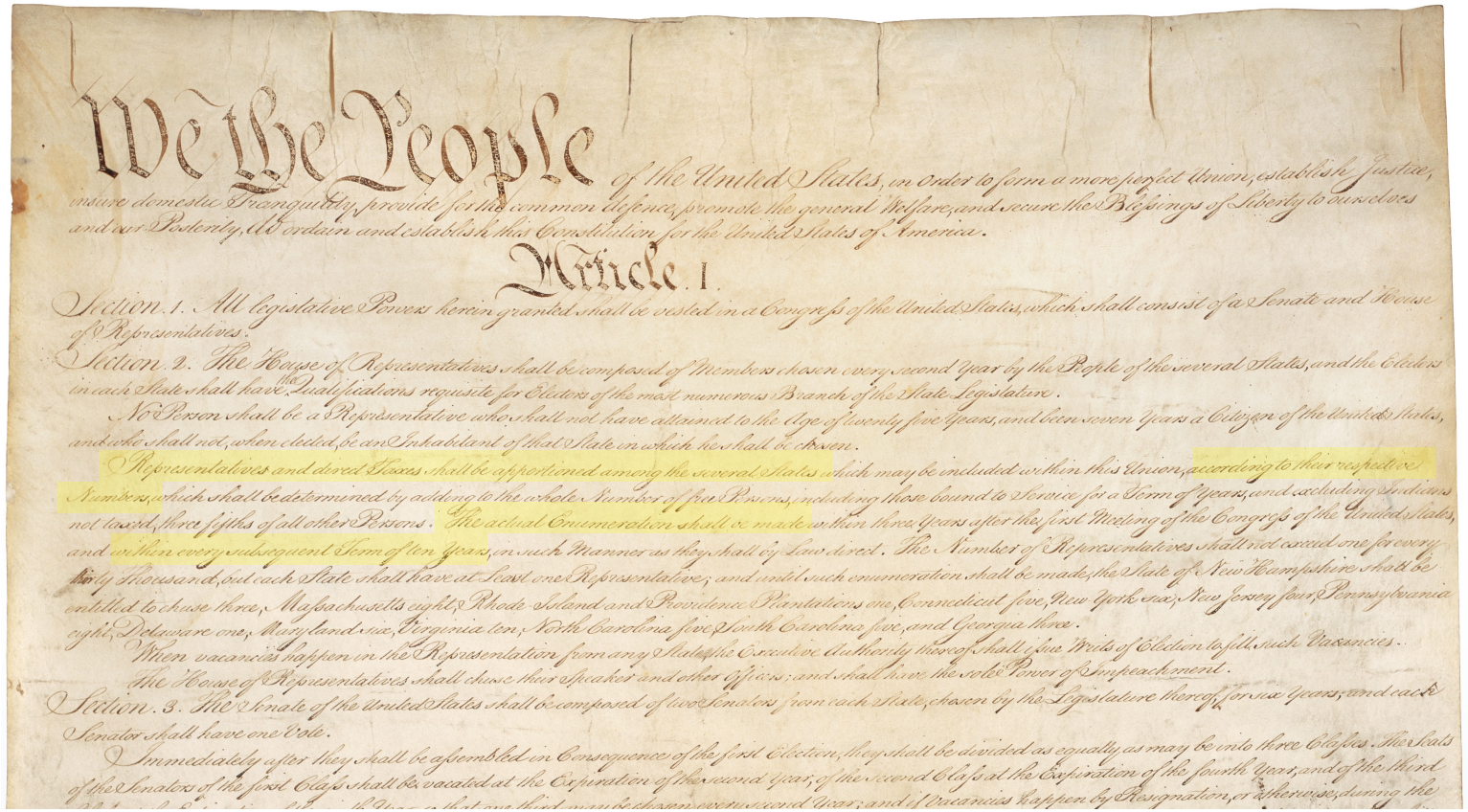


Do individual voters have standing?

- Some voters' power diluted, some amplified
- Catch 22: Can't know which is which.

Future directions

- More work (empirical + theory) needed on…
  - Reconstruction & privacy
  - Downstream policy impacts
  - ”Noisy measurement files”
- What’s next from Census?
  → Next talk!


We the People of the United States, in order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of America.

Article. I.

Section. 1. All legislative Powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.

Section. 2. The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in each State shall have the Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

No 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 States, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, 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. The Number of Representatives shall not exceed one for every thirty Thousand, but each State shall have at least one Representative; and until such enumeration shall be made, the State of New Hampshire shall be entitled to chuse three, Massachusetts eight, Rhode Island and Providence Plantations one, Connecticut five, New York six, New Jersey four, Pennsylvania eight, Delaware one, Maryland six, Virginia ten, North Carolina five, South Carolina five, and Georgia three.

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

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

Section. 3. The Senate of the United States shall be composed of two Senators from each State, chosen by the Legislature thereof, for six Years; and each Senator shall have one Vote.

Immediately after they shall be assembled in Consequence of the first Election, they shall be divided as equally as may be into three Classes. The Seats of the Senators of the first Class shall be vacated at the Expiration of the second Year, of the second Class at the Expiration of the fourth Year, and of the third