Intro to Differential Privacy

David Van Riper

vanriper@umn.edu

2020 Census demonstration data: Privacy and accuracy issues

December 4, 2019

Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau

Fri Aug 17 2018

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY

Protecting the Confidentiality of America's Statistics: Ensuring Confidentiality and Fitness-for-Use

Tue Sep 04 2018

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY

Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census

Fri Feb 15 2019

WRITTEN BY: DR. RON JARMIN, DEPUTY DIRECTOR AND COO

Census Bureau Continues to Boost Data Safeguards

Tue Jul 30 2019

WRITTEN BY: RON JARMIN, DEPUTY DIRECTOR, US CENSUS BUREAU

SUBSCRIBE | RSS SMS EMAIL

Outline

- What is differential privacy?
- Applying differential privacy to data
- Policy decisions
- Analyzing impact of differential privacy on 2010 decennial data

WHAT IS DIFFERENTIAL PRIVACY?

Differential privacy is…

- A formal (mathematical) definition of privacy

$$\frac{\Pr[M(D) \in S]}{\Pr[M(D') \in S]} \leq e^{\varepsilon}$$

Differential privacy is…

- A guarantee "on the incremental disclosure risks of participating in *D* over whatever disclosure risks the data subjects face even if they do not participate in *D*." (Reiter 2019)

Differential privacy is not…

- An algorithm for disclosure control

Differential privacy is not…

- An algorithm for disclosure control
- An absolute guarantee against disclosure risk

APPLYING DIFFERENTIAL PRIVACY

“True” microdata

Sex **School**

Male | Never
Male | Never
Male | Never

x12
Male | Attending
Male | Attending
⋮
Male | Attending

x33
Male | Past
⋮
Male | Past

Sex **School**

x4
Female | Never
⋮
Female | Never

x17
Female | Attending
⋮
Female | Attending

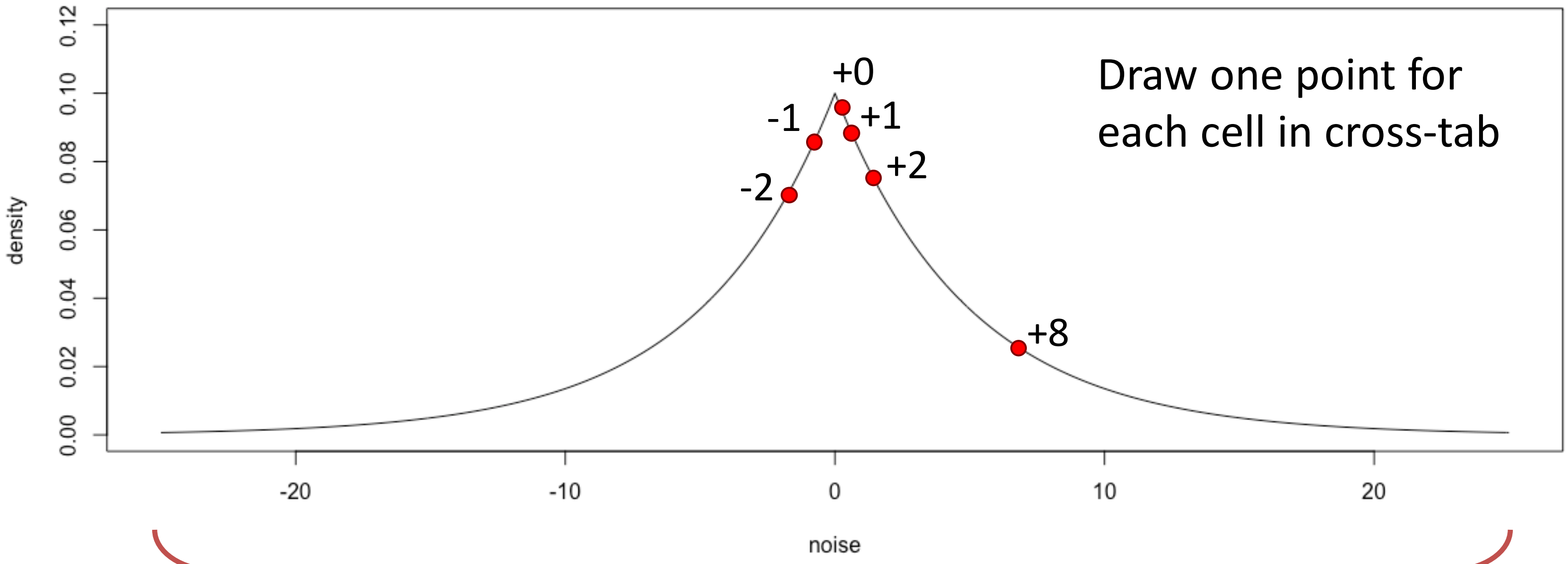
x31
Female | Past
⋮
Female | Past

Construct cross-tabs from “true” data

	School Attendance		
	Never	Attending	Past
Male	3	12	33
Female	4	17	31

Population = 100

Draw noise from Laplace distribution



spread is determined by **ε**

Add noise to cross-tab

	School Attendance		
	Never	Attending	Past
Male	3 – 1 = **2**	12 + 0 = **12**	33 + 1 = **34**
Female	4 + 8 = **12**	17 + 2 = **19**	31 – 2 = **29**

Sum = 108

Construct synthetic microdata

Male | Never
Male | Never

x12
Male | Attending
Male | Attending
⋮
Male | Attending

x34
Male | Past
⋮
Male | Past

x12
Female | Never
⋮
Female | Never

x19
Female | Attending
⋮
Female | Attending

x29
Female | Past
⋮
Female | Past

DIFFERENTIAL PRIVACY AND CENSUS

Differential privacy and census

POLICY DECISIONS

Policy decisions

- Global privacy loss budget (**ε**)

2010 demonstration data

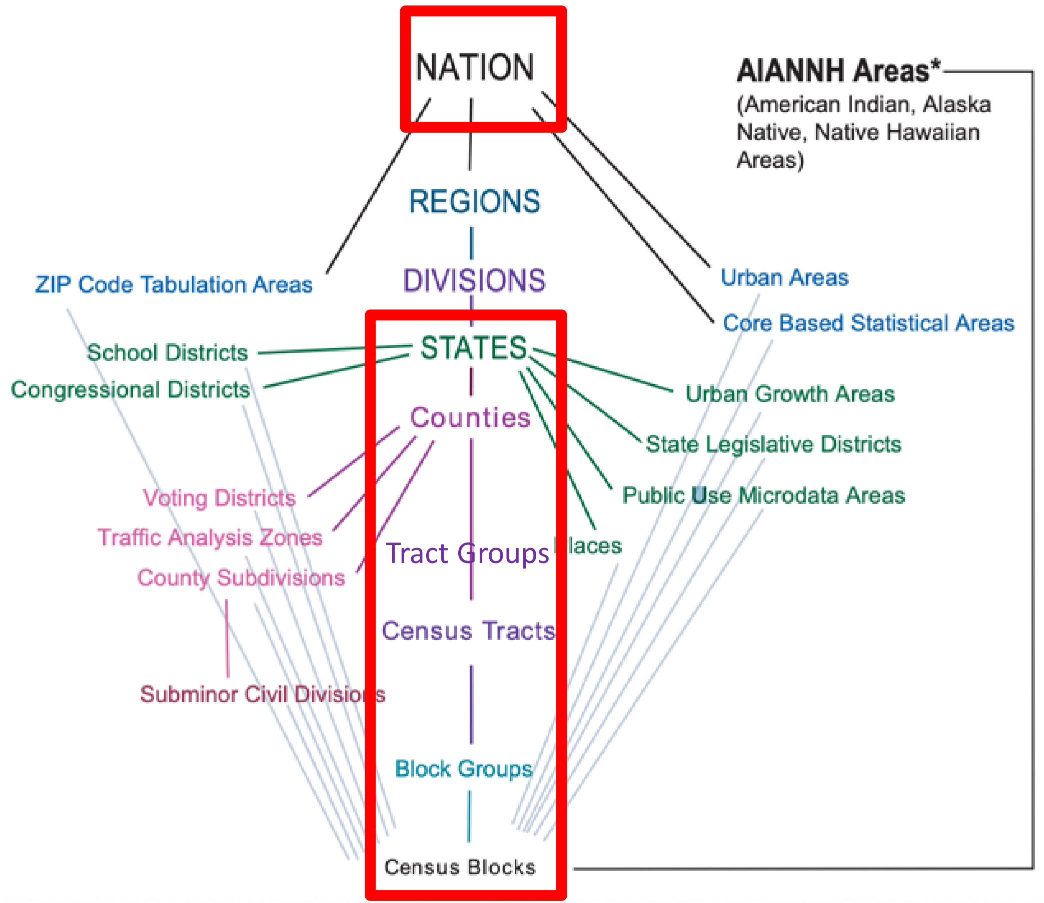
- Person tables
  - **ε** = 4.0
- Housing tables
  - **ε** = 2.0
- Global privacy loss budget
  - **ε** = 6.0

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level



NATION
AIANNH Areas*
(American Indian, Alaska Native, Native Hawaiian Areas)
REGIONS
DIVISIONS
ZIP Code Tabulation Areas
Urban Areas
Core Based Statistical Areas
STATES
School Districts
Congressional Districts
Urban Growth Areas
State Legislative Districts
Counties
Public Use Microdata Areas
Voting Districts
Traffic Analysis Zones
County Subdivisions
Tract Groups
laces
Census Tracts
Subminor Civil Divisions
Block Groups
Census Blocks

IPUMS.ORG

Geog_level	Fraction_{geog}
Nation	**0.2**
State	**0.2**
County	0.12
Tract Group	0.12
Tract	0.12
Block Group	0.12
Block	0.12

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level
- Tables

Policy decisions

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

- 2010 demonstration tables (examples)
  - Detailed person
    - Age x Sex x Hispanic x Race x HHGQ x Citizen
  - Voting age x Hispanic x Race x Citizen
  - Age x Sex
  - Detailed housing
  - Hispanic x Race x Size of HH x HH type

Person Tables	**$\text{Fraction}_{\text{table}}$**
Detailed	0.1
Household/Group Quarters Type	0.2
Voting age * Hispanic * Race * Citizen	**0.5**
Age * Sex	0.05
Age (4-year groups) * Sex	0.05
Age (16-year groups) * Sex	0.05
Age (64-year groups) * Sex	0.05

Housing Tables	**Fraction$_{table}$**
Detailed	0.2
Hispanic * Race * Size * HH_Type	**0.25**
HH_Sex * Hispanic * Race * HH_Type	**0.25**
Hisp * Race * Multi-generational	0.1
HH_Sex * HH_Type * Elderly	0.1
HH_Sex * HH_Age * HH_Type	0.1

Policy decisions

- Global privacy loss budget (**ε**)
- Geographic levels
  - Fraction of privacy budget allocated to each level
- Tables
  - Fraction of privacy budget allocated to each table
- Invariants and constraints

Policy decisions

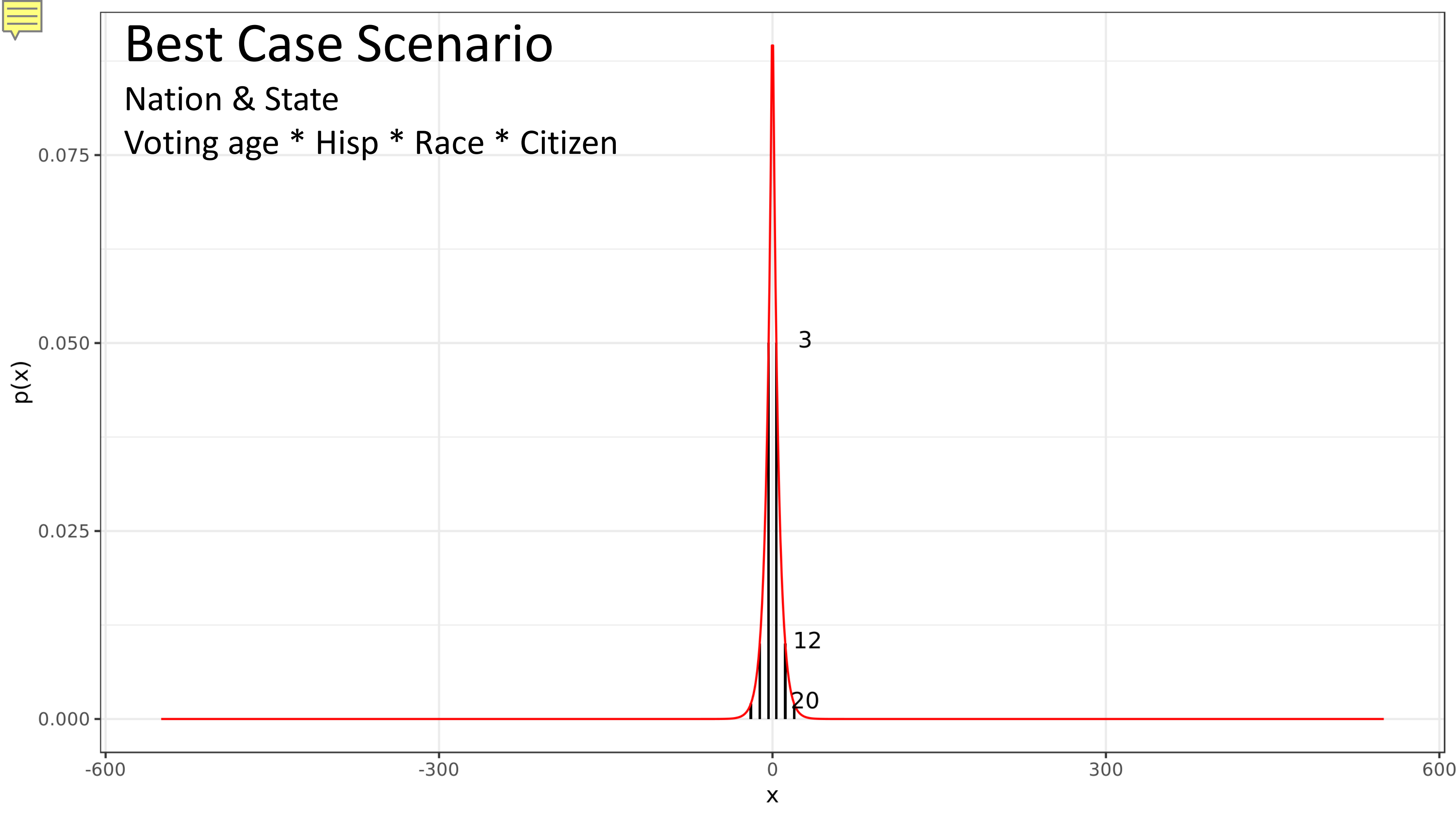
- Invariants (2010 demonstration data)
  - State-level total population
  - Census block-level total housing units
  - Census block-level group quarters count
  - Census block-level group quarters type count

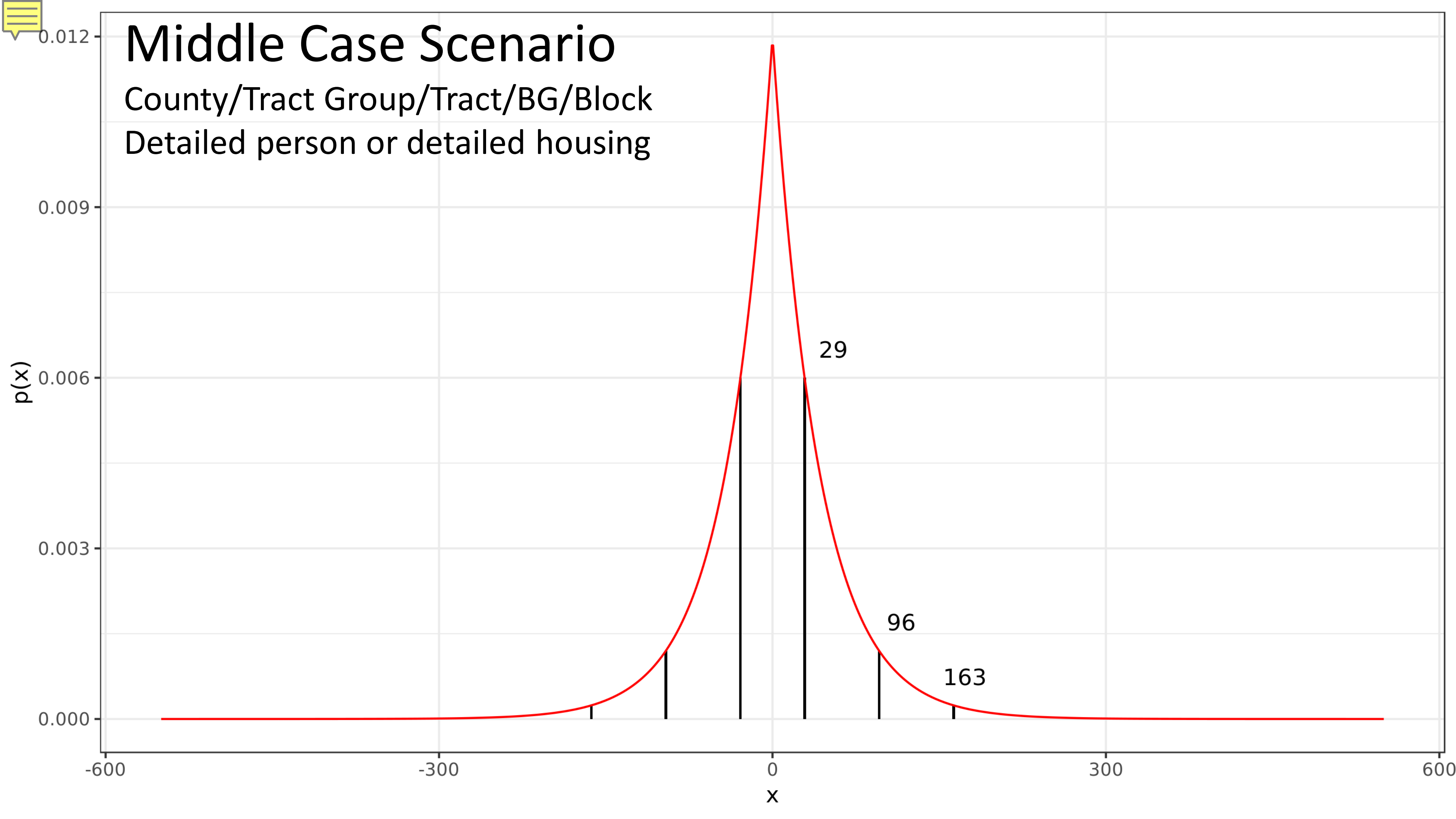
Policy decisions

- Invariants (2010 decennial data)
  - Census block-level total population
  - Census block-level voting age population
  - Census block-level total housing units
  - Census block-level occupancy status
  - Census block-level group quarters count
  - Census block-level group quarters type count

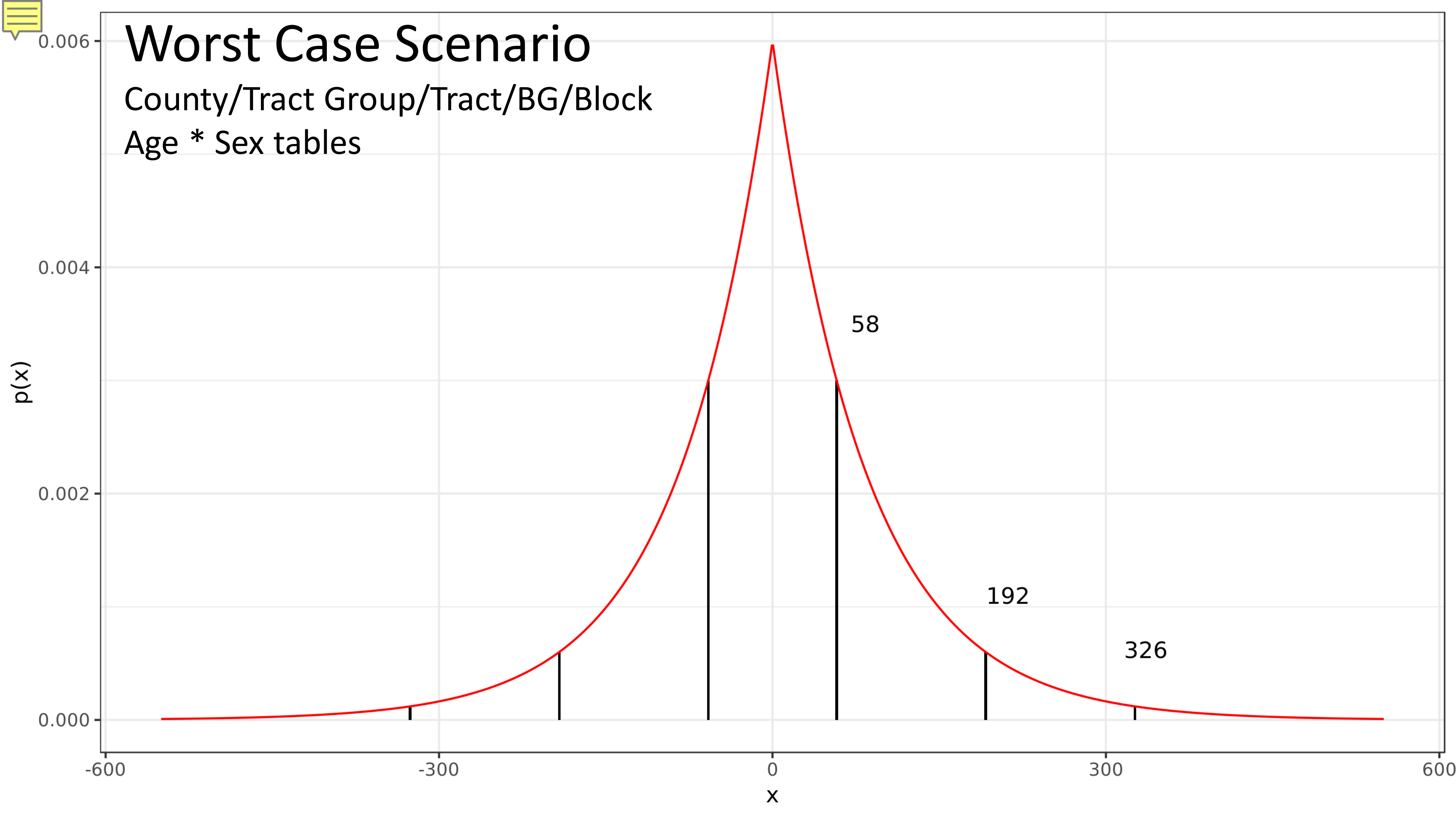
Technical Implementation

NOISE INJECTION


Best Case Scenario
Nation & State
Voting age * Hisp * Race * Citizen
0.075
0.050
0.025
0.000
p(x)
3
12
20
-600
-300
0
300
600
x



Middle Case Scenario
County/Tract Group/Tract/BG/Block
Detailed person or detailed housing
29
96
163
0.012
0.009
0.006
0.003
0.000
p(x)
-600
-300
0
300
600
x


Worst Case Scenario
County/Tract Group/Tract/BG/Block
Age * Sex tables
58
192
326
0.006
0.004
0.002
0.000
p(x)
-600
-300
0
300
600
x

FULL IMPLEMENTATION

1. Generate microdata without geographic identifiers
2. Assign geographic identifiers to each microdata record

Step 1

1. Create national tables from “true” data

Step 1

1. Create national tables from “true” data
2. For each cell in tables, infuse noise drawn from Laplace/geometric distribution

Step 1

1. Create national tables from “true” data
2. For each cell in tables, infuse noise drawn from Laplace/geometric distribution
3. Generate microdata with no geographic identifiers from (2) via database reconstruction

Step 2

1. Create state tables from “true” data

Step 2

1. Create state tables from “true” data
2. For each cell in tables, infuse noise drawn from Laplace /geometric distribution

Step 2

1. Create state tables from “true” data
2. For each cell in tables, infuse noise drawn from Laplace /geometric distribution
3. Use linear optimization to fit Step 1 microdata to “noisy” state cells

Step 2

1. Create state tables from “true” data
2. For each cell in tables, infuse noise drawn from Laplace /geometric distribution
3. Use linear optimization to fit Step 1 microdata to “noisy” state cells
4. Assign state identifier to each Step 1 microdata record

Step 2

1. Create state tables from “true” data
2. For each cell in tables, infuse noise drawn from Laplace /geometric distribution
3. Use linear optimization to fit Step 1 microdata to “noisy” state cells
4. Assign state identifier to each Step 1 microdata record
5. Repeat (1) – (4) for remaining geographic levels (counties down to census blocks)

Output

- Microdata records with state, county, tract group, census tract, census block group, and census block identifiers

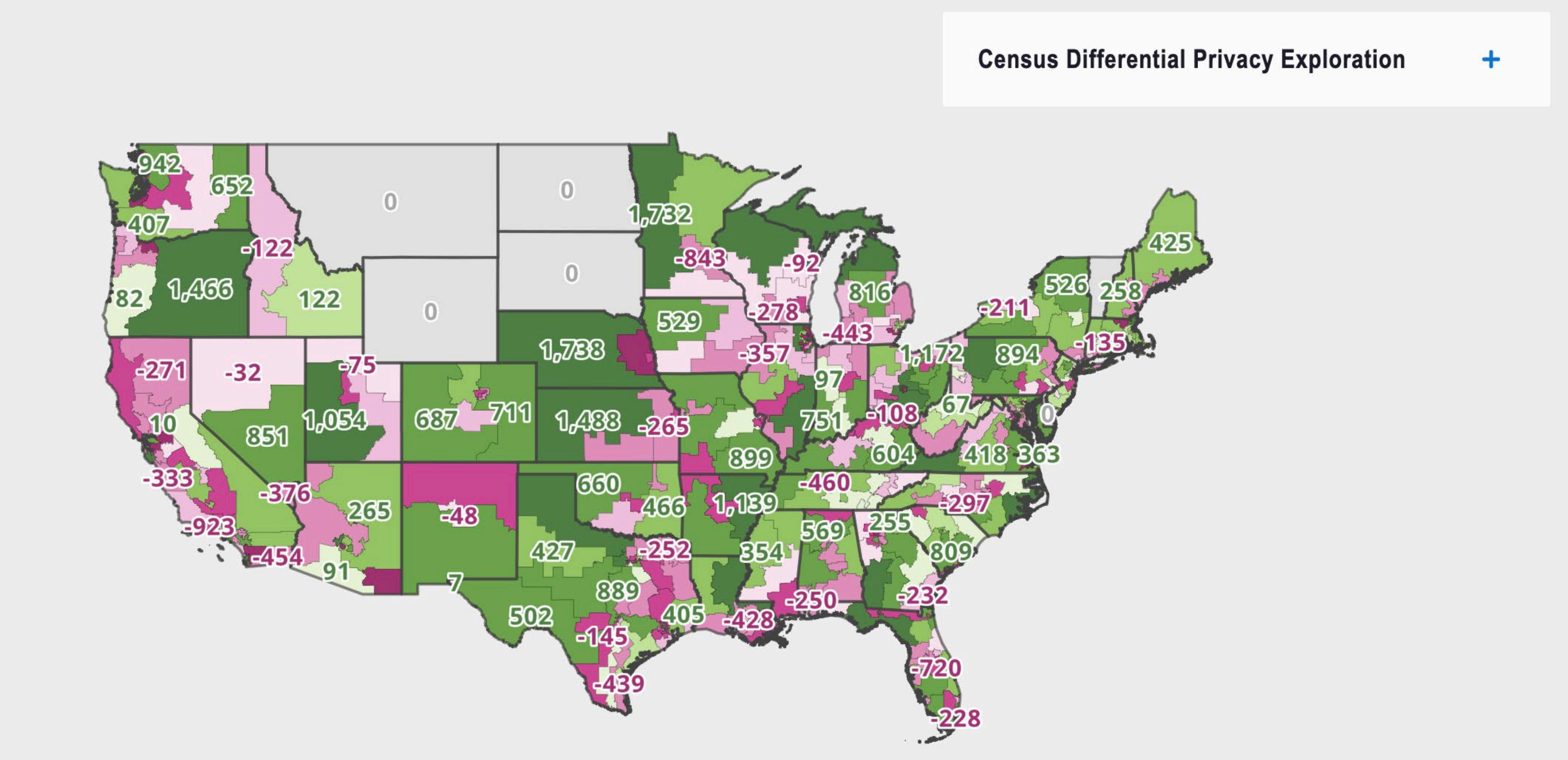
ANALYZING DIFFERENTIALLY PRIVATE 2010 CENSUS DATA

Table 3. Percent difference greater than 10% or 25% for total population.

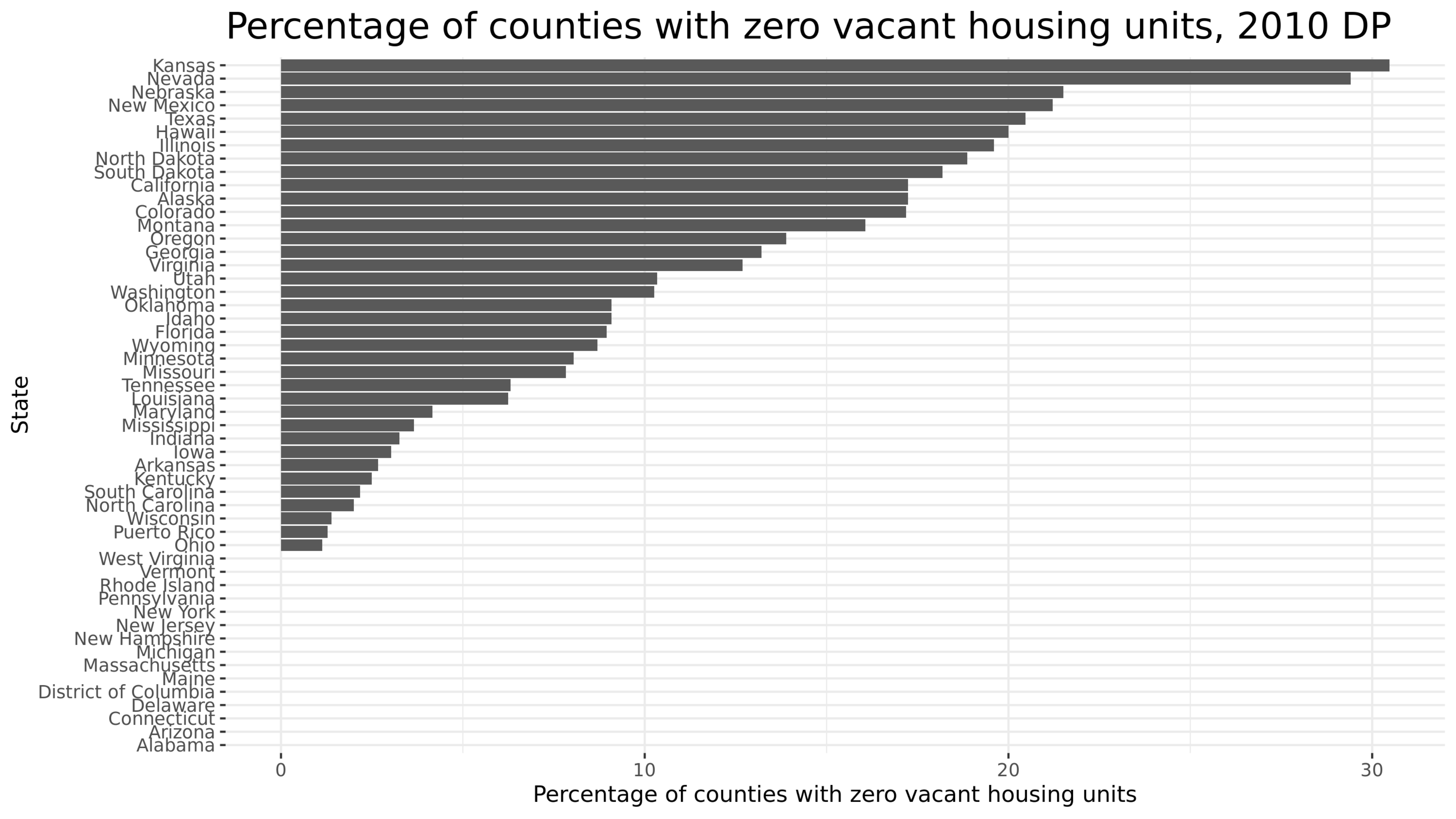
Geographic level	Units	10% or more	25% or more
Counties	3,221	13	2
County subdivisions	36,642	8,599	3,958
Places	29,514	8,109	3,345
Urban areas	3,592	31	0
AIANHH	692	326	195

Table 4. Percent difference in total population for place deciles.

Decile	Mean total population (SF1)	Mean percent difference	10% or more	25% or more
1	73.9	38.2	2,390	1,640
2	190.4	19.6	1,907	890
3	336.7	14.7	1,617	505
4	548.7	10.2	1,164	212
5	866.7	6.6	636	79
6	1,385.7	4.4	291	16
7	2,291.9	2.9	88	3
8	4,139.0	1.9	16	0
9	8,939.2	1.2	0	0
10	59,353.0	0.6	0	0

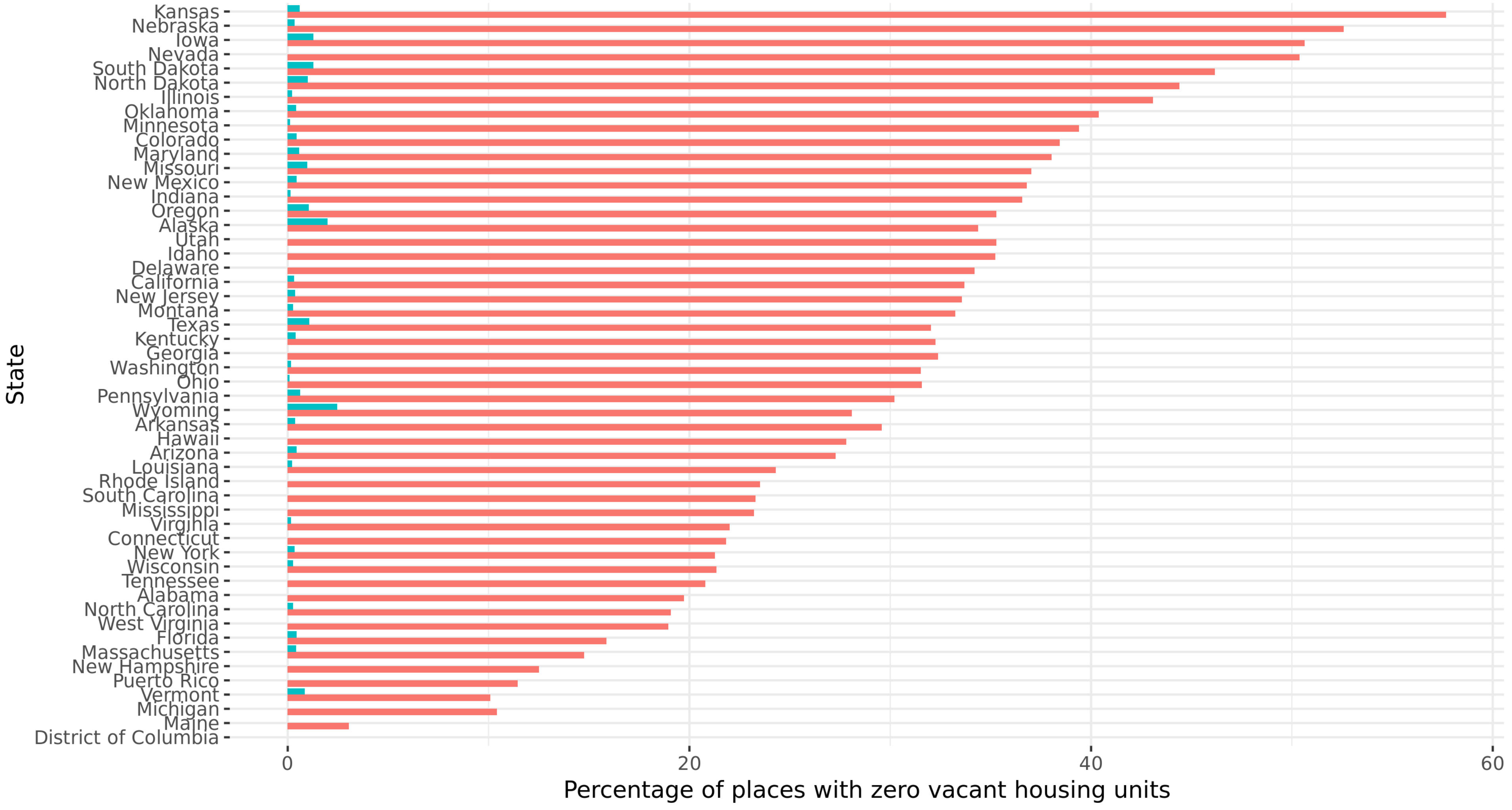


Source: https://www.caliper.com/census-differential-privacy-maps/



Percentage of counties with zero vacant housing units, 2010 DP
State
Kansas
Nevada
Nebraska
New Mexico
Texas
Hawaii
Illinois
North Dakota
South Dakota
California
Alaska
Colorado
Montana
Oregon
Georgia
Virginia
Utah
Washington
Oklahoma
Idaho
Florida
Wyoming
Minnesota
Missouri
Tennessee
Louisiana
Maryland
Mississippi
Indiana
Iowa
Arkansas
Kentucky
South Carolina
North Carolina
Wisconsin
Puerto Rico
Ohio
West Virginia
Vermont
Rhode Island
Pennsylvania
New York
New Jersey
New Hampshire
Michigan
Massachusetts
Maine
District of Columbia
Delaware
Connecticut
Arizona
Alabama
0
10
20
30
Percentage of counties with zero vacant housing units

Figure 2. Percentage of places with zero vacant housing units in 2010 DP and SF1 data.



Conclusions

- Diff. privacy less complicated than expected

Conclusions

- Diff. privacy less complicated than expected
- Fundamental importance of policy decisions

Conclusions

- Diff. privacy less complicated than expected
- Fundamental importance of policy decisions
- Largest impact on accuracy of small areas and small sub-populations

References

2020 Census DAS Development Team. (2019) 2019. *Disclosure Avoidance System for the 2020 Census, End-to-End Release: Uscensusbureau/Census2020-Das-E2e*. Python. US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e.

Abowd, John. 2018. "Disclosure Avoidance for Block Level Data and Protection of Confidentiality in Public Tabulations." US Census Bureau. https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#.

Abowd, John, Daniel Kifer, Brett Moran, Robert Ashmead, Philip Leclerc, William Sexton, Simson Garfinkel, and Ashwin Machanavajjhala. 2019. "Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge." US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf.

Ashmead, Robert, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. 2019. "Effective Privacy after Adjusting for Invariants with Applications to the 2020 Census." US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0941_Effective_Privacy_after_Adjusting_for_Constraints__With_applications_to_the_2020_Census.pdf.

Bambauer, Jane, Krishnamurty Muralidhar, and Rathindra Sarathy. n.d. "Fool's Gold: An Illustrated Critique of Differential Privacy." *Vanderbilt Journal of Entertainment & Technology Law* 16: 55.

boyd, danah. 2019. "Differential Privacy in the 2020 Decennial Census and the Implications for Available Data Products." *ArXiv:1907.03639 [Cs]*, July. http://arxiv.org/abs/1907.03639.

Census Bureau, US. 2019. “2018 End-to-End Test Disclosure Avoidance System Design Specification, Version 1.2.8.” US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/2019-04-11-2018-End-to-End-Test-Disclosure-Avoidance-System-Design-Specification.pdf.

———. n.d. “Differentially Private 1940 Census Data.” US Census Bureau. https://www2.census.gov/census_1940/.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. “Issues Encountered Deploying Differential Privacy.” In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–137. WPES’18. New York, NY, USA: ACM. https://doi.org/10.1145/3267323.3268949.

Iceland, John. 2004. “The Multigroup Entropy Index (Also Known as Theil’s H or the Information Theory Index).” US Census Bureau. https://www2.census.gov/programs-surveys/demo/about/housing-patterns/multigroup_entropy.pdf.

Leclerc, Philip. 2019. “Guide to the Census 2018 End-to-End Test Disclosure Avoidance Algorithm and Implementation.” US Census Bureau. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0938_2018_E2E_Test_Algorithm_Description.pdf.

McClure, David, and Jerome P. Reiter. 2012. “Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data.” *Trans. Data Privacy* 5 (3): 535–552.

Reiter, Jerome P. 2019. “Differential Privacy and Federal Data Releases.” *Annual Review of Statistics and Its Application* 6 (1): 85–101. https://doi.org/10.1146/annurev-statistics-030718-105142.

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. “Differential Privacy and Census Data: Implications for Social and Economic Research.” *AEA Papers and Proceedings* 109 (May): 403–8. https://doi.org/10.1257/pandp.20191107.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2018. "IPUMS USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research [Dataset]." IPUMS. https://doi.org/10.18128/D010.V8.0.EXT1940USCB.

Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. 2018. "Differential Privacy: A Primer for a Non-Technical Audience." *Vanderbilt Journal of Entertainment & Technology Law*, no. 1 (2019): 209–76.

- 2010 demonstration tables (examples)
  - Voting age [2] x Hispanic [2] x Race [63] x Citizen [2]
    - 2 * 2 * 63 * 2 = 504 cells
  - Age [116] x Sex [2] x Race [63] x Hispanic [2] x HHGQ [8] x Citizen [2]
    - 116 * 2 * 63 * 2 * 8 * 2 = 467,712 cells
  - Age [116] x Sex [2]
    - 116 * 2 = 232 cells

Geog_level
Nation
State
County
Tract Group
Tract
Block Group
Block

Geog_level	$\text{Fraction}_{\text{geog}}$
Nation	0.2
State	0.2
County	0.12
Tract Group	0.12
Tract	0.12
Block Group	0.12
Block	0.12

Geog_level	**$\text{Fraction}_{\text{geog}}$**
Nation	**0.2**
State	**0.2**
County	0.12
Tract Group	0.12
Tract	0.12
Block Group	0.12
Block	0.12

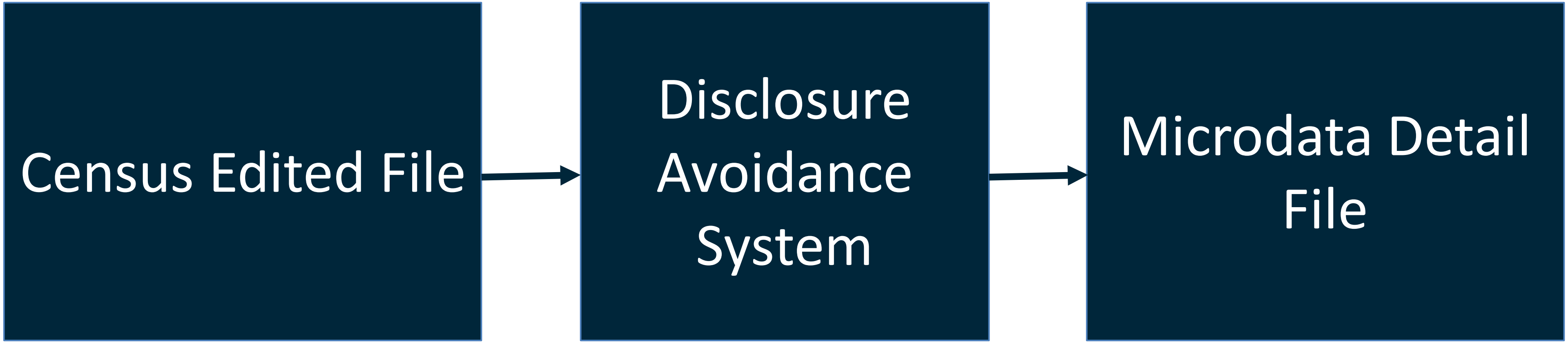
Geog_level
Nation
State
County
Tract Group
Tract
Block Group
Block

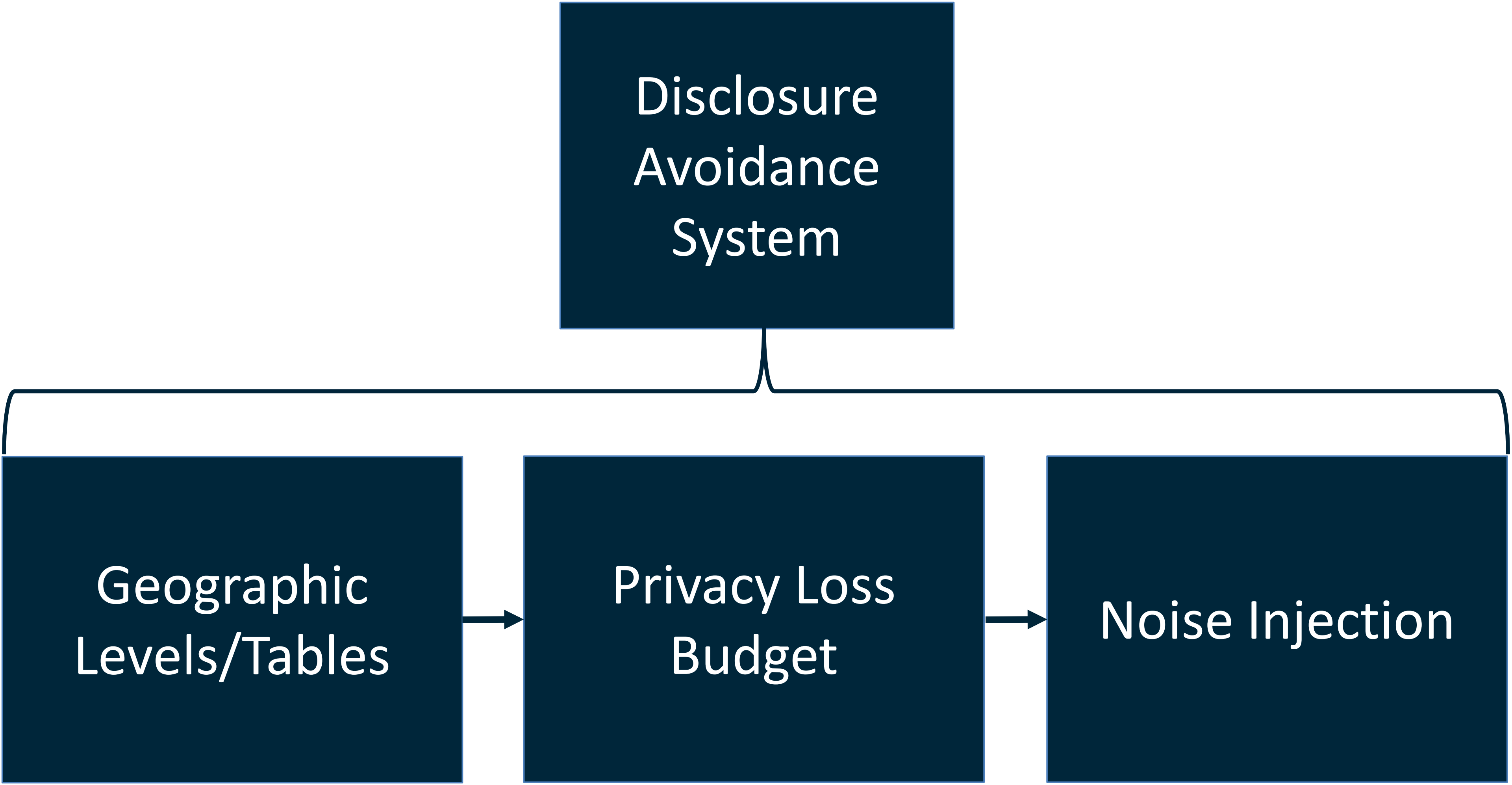
X

Table
Detailed
HHGQ
Voting age * Hispanic * Race * Citizen
Age * Sex
Age (4 year groups) * Sex
Age (16 year groups) * Sex
Age (64 year groups) * Sex

Differential privacy and census

TECHNICAL IMPLEMENTATION


Census Edited File
Disclosure Avoidance System
Microdata Detail File


Disclosure Avoidance System
Geographic Levels/Tables
Privacy Loss Budget
Noise Injection