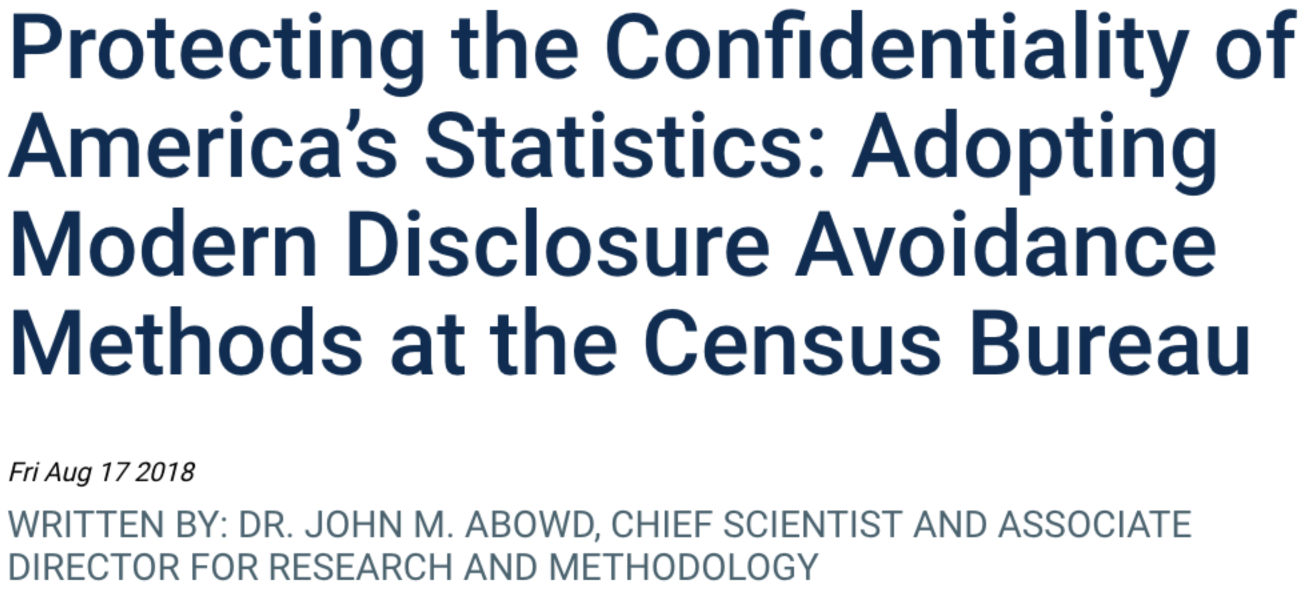
Disclosure Avoidance and the 2020 Decennial Census

David Van Riper

vanriper@umn.edu

@dcvanriper

User Experience Accessing Disaggregated Racial/Ethnic Data
National Network of Health Survey's Data Disaggregation Workshop
November 18, 2020



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

Outline

- How is differential privacy implemented?
- How does this new disclosure avoidance technique impact public health analyses?

HOW IS DIFFERENTIAL PRIVACY IMPLEMENTED?

“True” microdata

Sex **School**

Male | Never
Male | Never
Male | Never
x12
Male | Attending
Male | Attending
⋮
Male | Attending
x33
Male | Past
⋮
Male | Past

Sex **School**

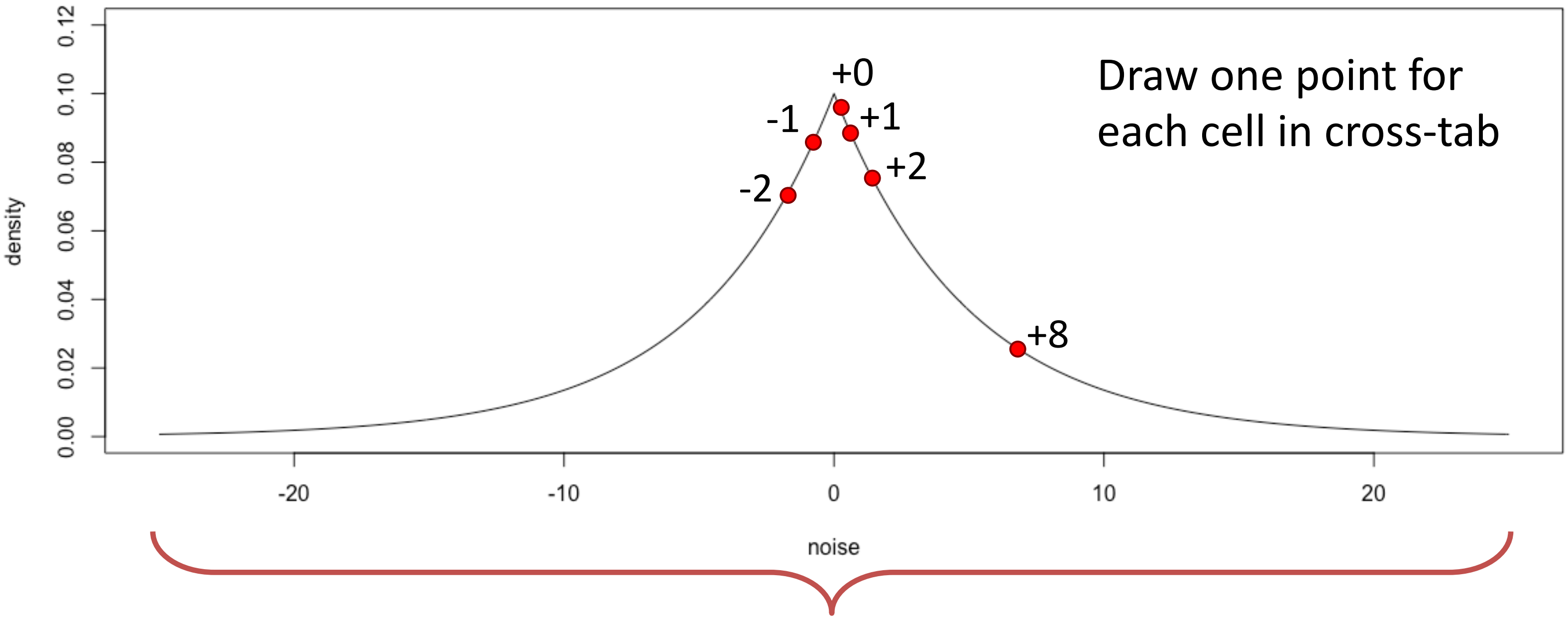
x4
Female | Never
⋮
Female | Never
x17
Female | Attending
⋮
Female | Attending
x31
Female | Past
⋮
Female | Past

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



Draw one point for each cell in cross-tab

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

POLICY DECISIONS

Policy decisions

- Global privacy loss budget (**ε**)
- Fractional allocations
- Invariants and constraints
- Post-processing

Global privacy loss budget

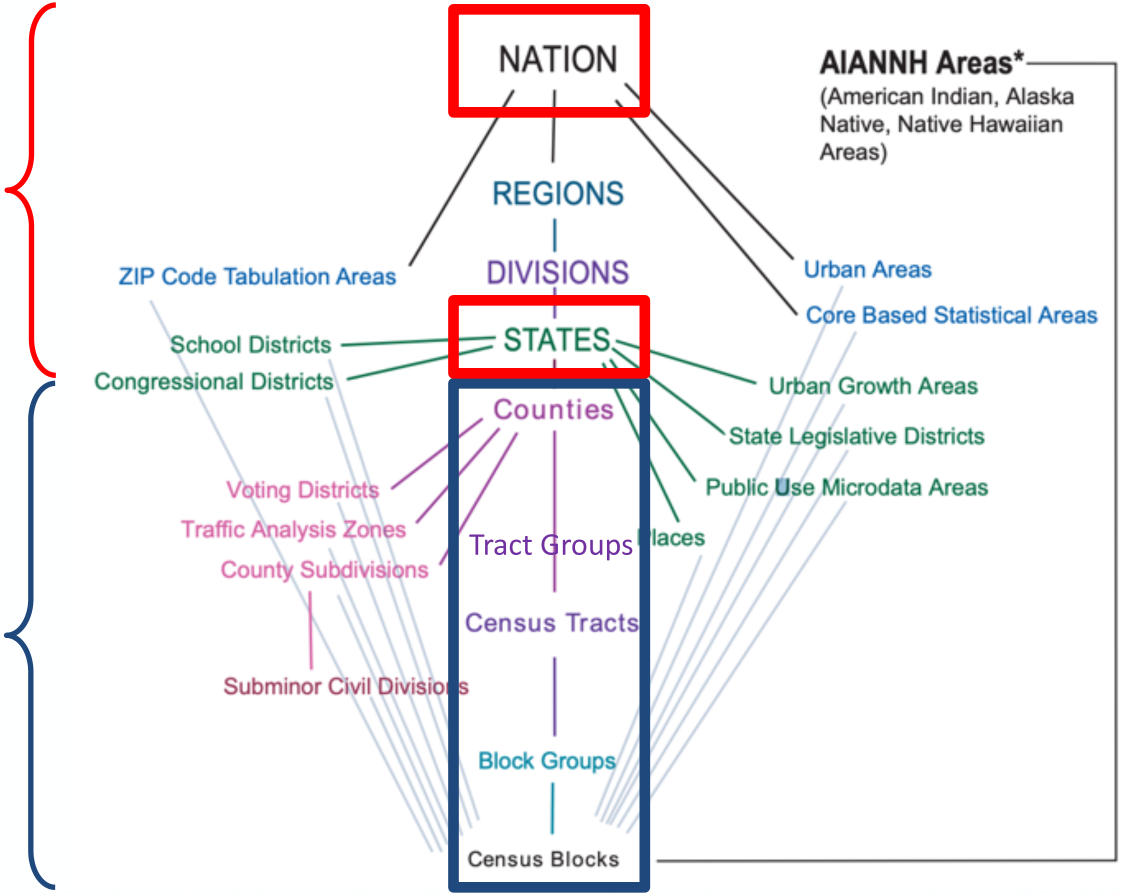
- Global privacy loss budget
  - **ε** = 6.0
- Person tables
  - **ε** = 4.0
- Housing tables
  - **ε** = 2.0

Fractional allocations

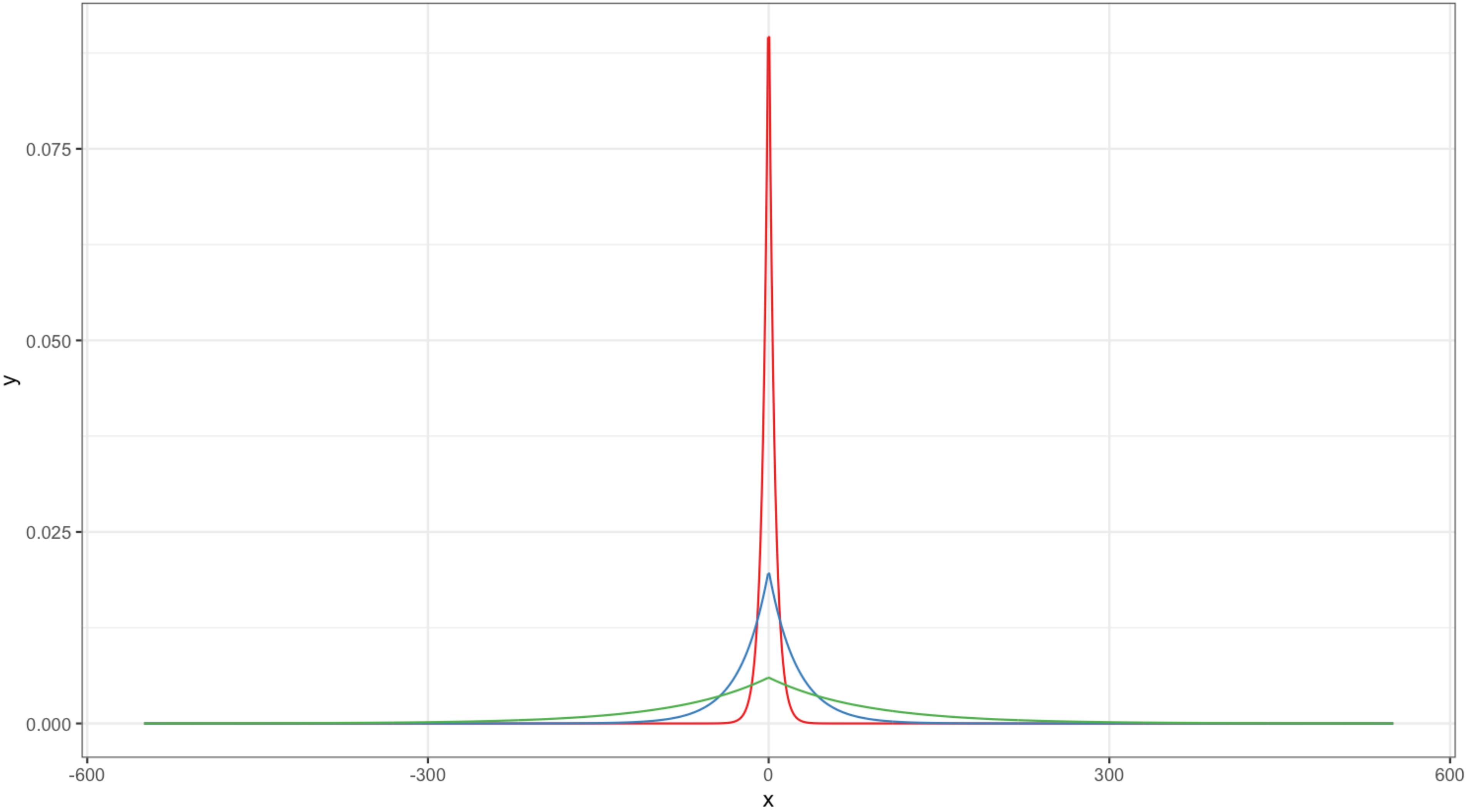
- Geographic levels
- Queries

20% each

12% each



Query	**Allocation (%)**
Voting age * Hispanic * Race * Citizen	50
Household – Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5


y
0.075
0.050
0.025
0.000
-600
-300
0
300
600
x

Invariants and Constraints

- Invariants are counts not subject to noise injection

2010 Decennial Invariants	2010 Demonstration Data Invariants
Total population (block)	Total population (state)
Total housing units (block)	Total housing units (block)
Group quarters count (block)	Group quarters count (block)
Group quarters type count (block)	Group quarters type count (block)
Occupancy status (block)	
Voting age population (block)	

2010 Decennial Invariants	2010 Demonstration Data Invariants
Total population (block)	Total population (state)
Total housing units (block)	Total housing units (block)
Group quarters count (block)	Group quarters count (block)
Group quarters type count (block)	Group quarters type count (block)
Occupancy status (block)	
Voting age population (block)	

Invariants and Constraints

- Invariants are counts not subject to noise injection
- Constraints

Invariants and Constraints

- Invariants are counts not subject to noise injection
- Constraints
  - Non-negativity
  - Consistency

Post-processing

- Non-negative least squares + constraints = positive bias for small counts and negative bias for large counts

ANALYZING DIFFERENTIALLY PRIVATE 2010 CENSUS DATA

Data

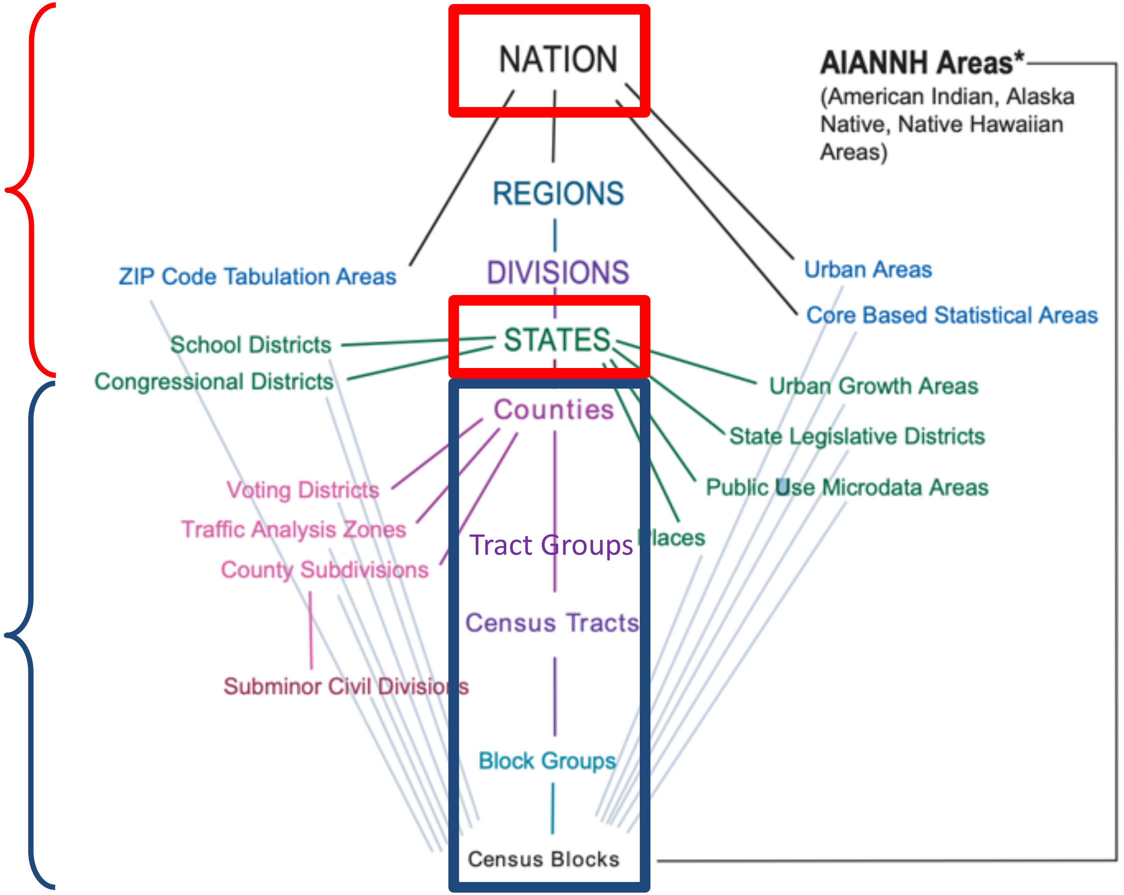
- 2010 Summary File 1
- Vintage 1 (October 2019)
- Vintage 2 (June 2020)

Comparisons

- Comparing data from vintage 1 and 2 with data from Summary File 1
- Summary File 1 essentially serves as our "ground truth"
    - Acknowledging that prior disclosure avoidance techniques introduced error into SF1

20% each

12% each



Vintage 1

Query	Allocation (%)
Voting age * Hispanic * Race * Citizen	50
Relation to HH/Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5

Vintage 2

Query	Allocation (%)
Total population	30
Voting age * Hispanic * Race	29
Age * Sex * Hispanic * Race	25
Relation to HH/Group quarters	15
Detailed	1

Vintage 1

Query	Allocation (%)
Voting age * Hispanic * Race * Citizen	50
Relation to HH/Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5

Vintage 2

Query	Allocation (%)
Total population	30
Voting age * Hispanic * Race	29
Age * Sex * Hispanic * Race	25
Relation to HH/Group quarters	15
Detailed	1

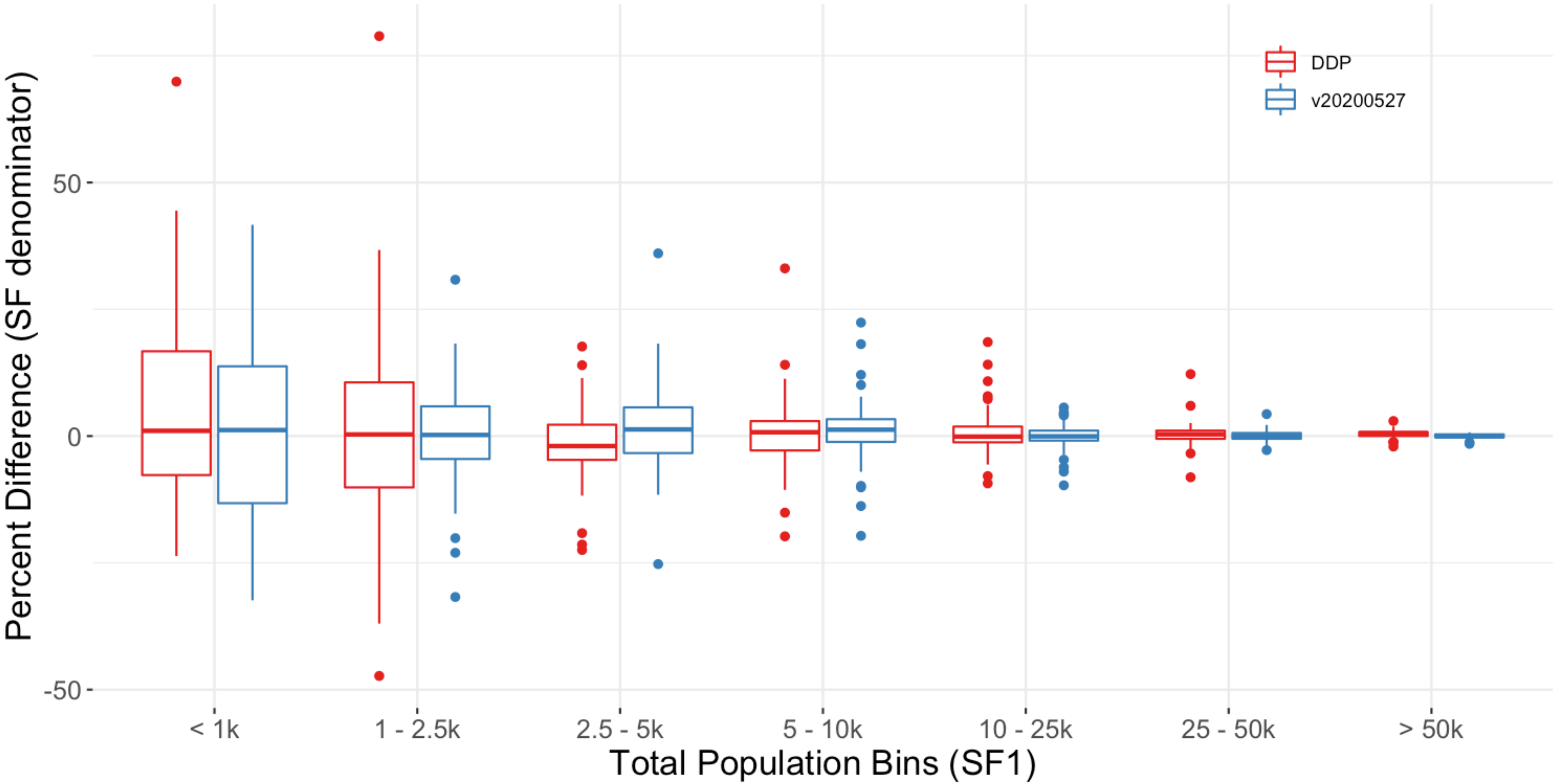
Age-adjusted rates of

- Asthma ED visits in 2010
  - Towns in Massachusetts
  - Counties in 25 states

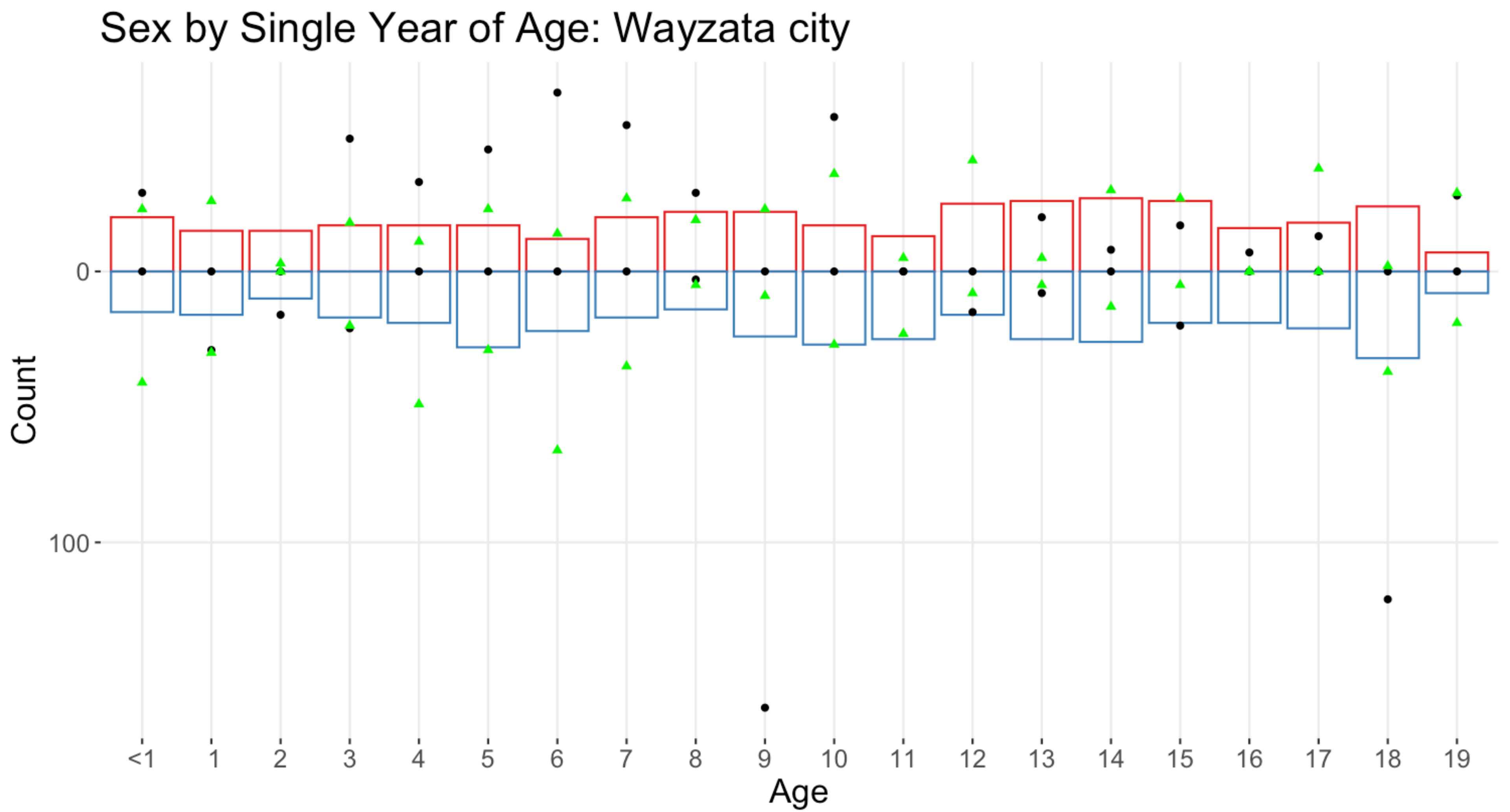
Rate comparison

$$PercentDifference = \frac{DP_{rate} - SF1_{rate}}{SF1_{rate}} * 100$$

Percent Difference in Age-Adjusted Asthma ED Visits in 2010 (MA towns)

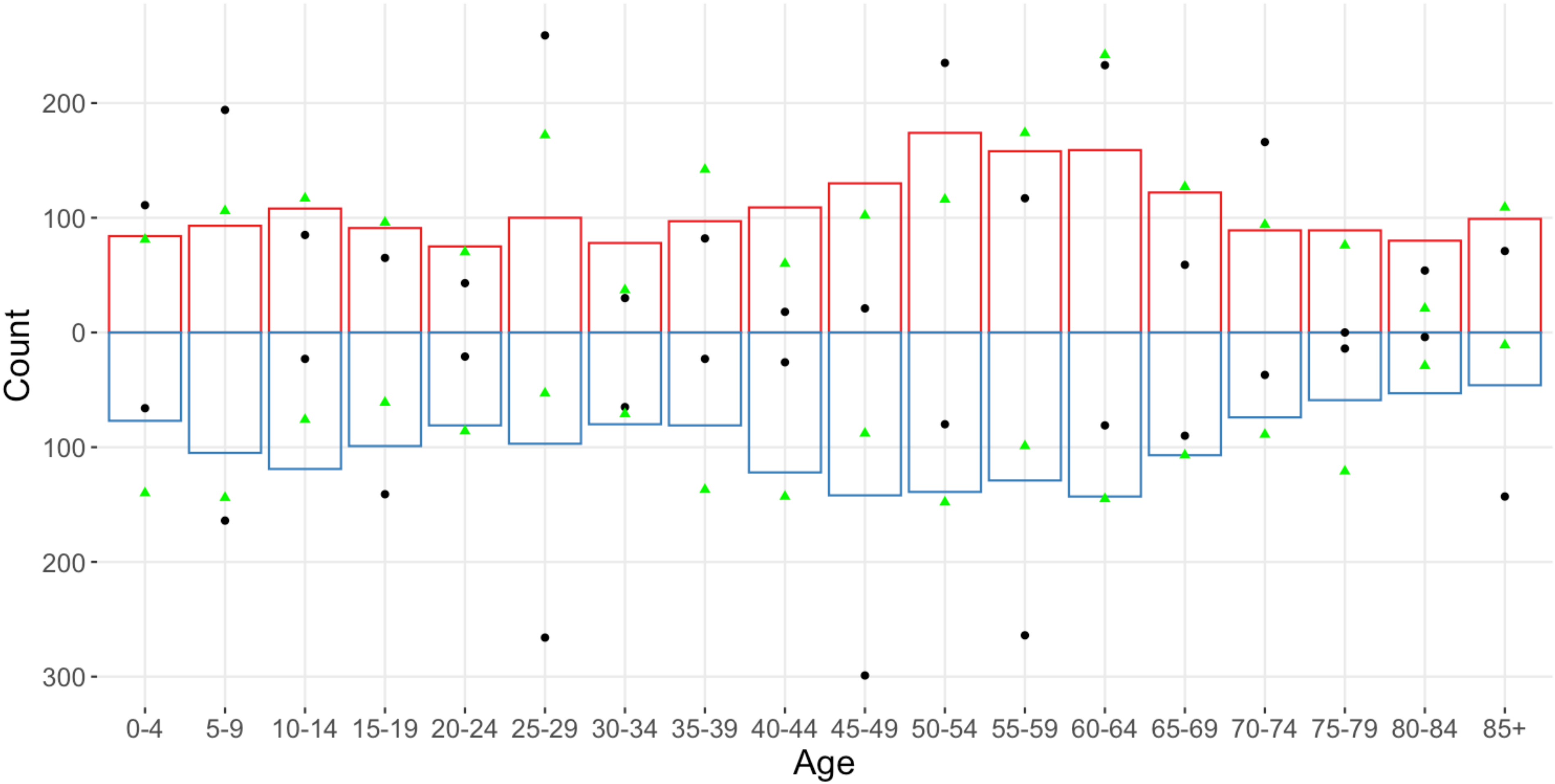


Source: Van Riper et al. 2020; US Census Bureau 2019; Massachusetts Department of Health 2020



Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

Sex by Age: G270053068818



Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

Conclusions

- Moving target – Census continuously changing disclosure avoidance algorithm
- Public health analysis will be impacted
  - subpopulations with small counts
  - geographic units with small counts
- Quantifying uncertainty important

Contact

- David Van Riper
  - vanriper@umn.edu
- Differentially private summary data
  - DDP
    - https://www.nhgis.org/differentially-private-2010-census-data
  - V20200527
    - https://nhgis.org/privacy-protected-demonstration-data