Differential Privacy

Jan Vink

Why can Census data not be fully released?

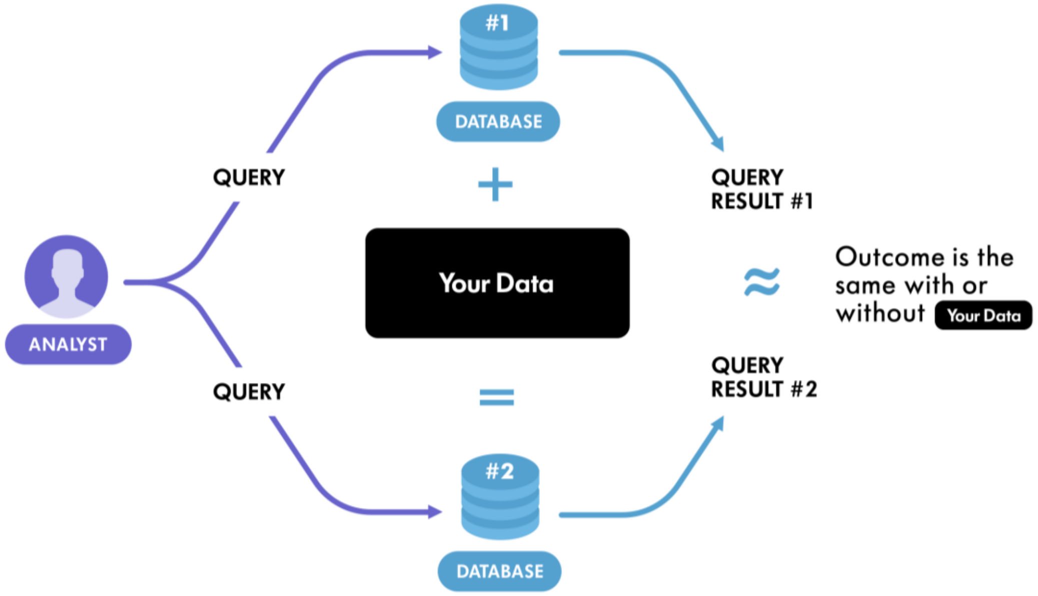
- Under Title 13 of the U.S. Code, the Census Bureau cannot release any identifiable information about individuals, households, or businesses, even to law enforcement agencies. The law states that the information collected may only be used for statistical purposes and no other purpose.

Possible solutions

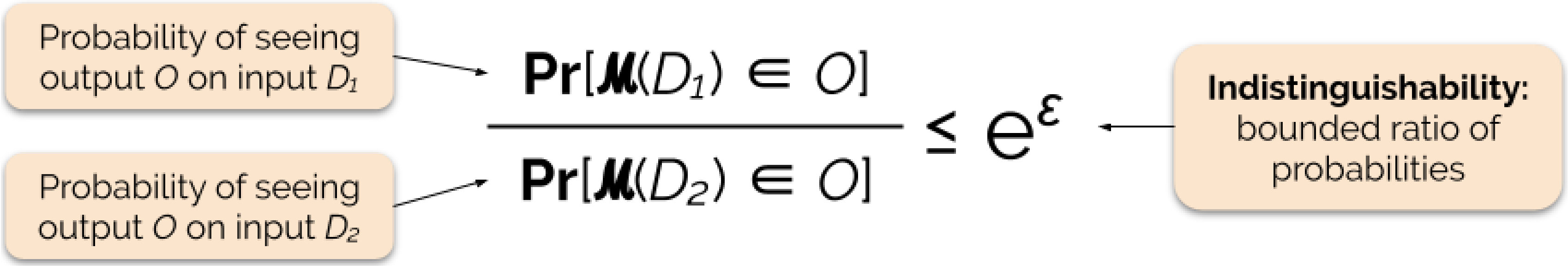
- Aggregate
- Suppression
- Adding uncertainty to the numbers
  - Rounding
  - Swapping
  - Differential Privacy

What is Differential Privacy

- Differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual’s private information, whether
  or not that individual’s private information is included
  in the input to the analysis.



What is Differential Privacy



Probability of seeing output *O* on input D_1

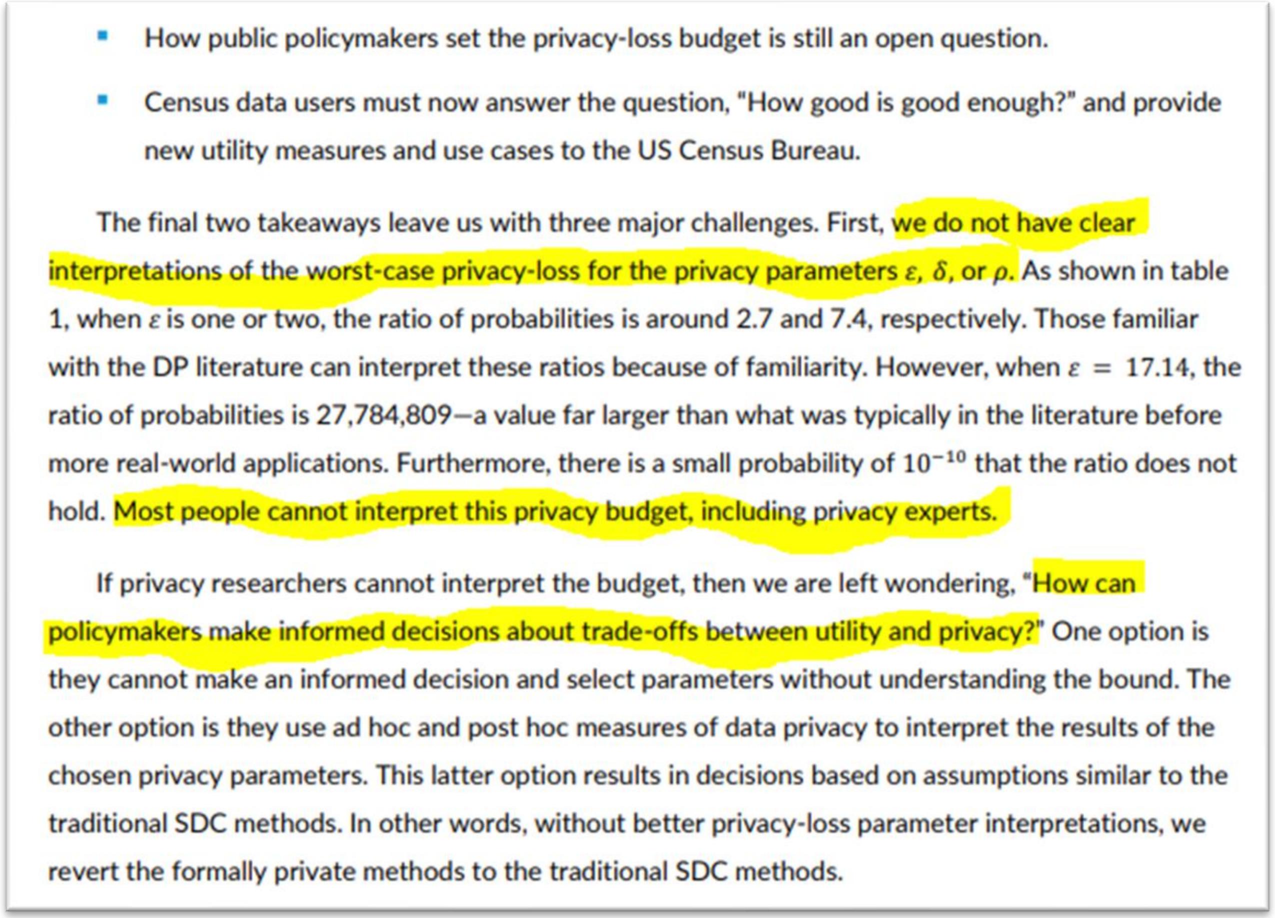
Probability of seeing output *O* on input D_2

$$\frac{\Pr[\mathcal{M}(D_1) \in O]}{\Pr[\mathcal{M}(D_2) \in O]} \leq e^{\varepsilon}$$

Indistinguishability: bounded ratio of probabilities

- The strength of this privacy guarantee is a policy decision and is often expressed as Epsilon (lower epsilon is more privacy)
- The higher the privacy guarantee the more the data need to be ‘fudged’ and the more risk that usability suffers

From a recent DP explainer



- How public policymakers set the privacy-loss budget is still an open question.
- Census data users must now answer the question, "How good is good enough?" and provide new utility measures and use cases to the US Census Bureau.

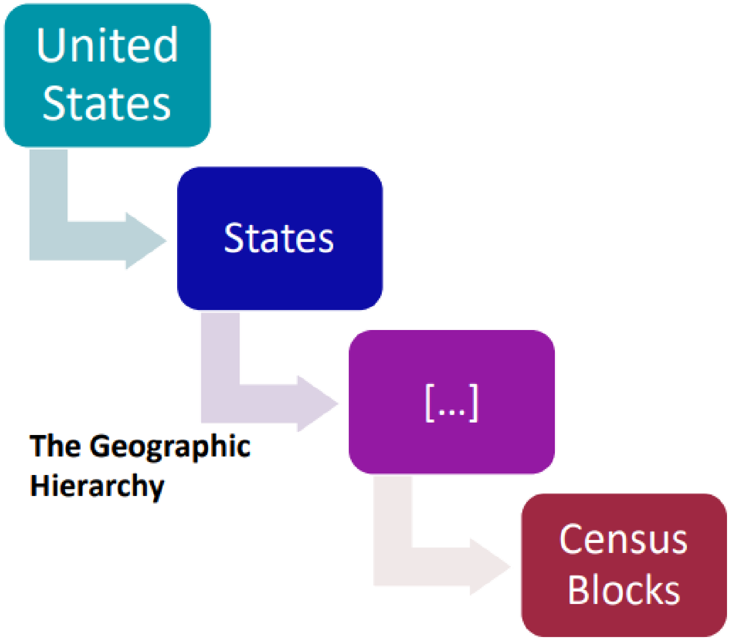
The final two takeaways leave us with three major challenges. First, we do not have clear interpretations of the worst-case privacy-loss for the privacy parameters ε, δ, or ρ. As shown in table 1, when ε is one or two, the ratio of probabilities is around 2.7 and 7.4, respectively. Those familiar with the DP literature can interpret these ratios because of familiarity. However, when $\varepsilon = 17.14$, the ratio of probabilities is 27,784,809—a value far larger than what was typically in the literature before more real-world applications. Furthermore, there is a small probability of 10^{-10} that the ratio does not hold. Most people cannot interpret this privacy budget, including privacy experts.

If privacy researchers cannot interpret the budget, then we are left wondering, "How can policymakers make informed decisions about trade-offs between utility and privacy?" One option is they cannot make an informed decision and select parameters without understanding the bound. The other option is they use ad hoc and post hoc measures of data privacy to interpret the results of the chosen privacy parameters. This latter option results in decisions based on assumptions similar to the traditional SDC methods. In other words, without better privacy-loss parameter interpretations, we revert the formally private methods to the traditional SDC methods.

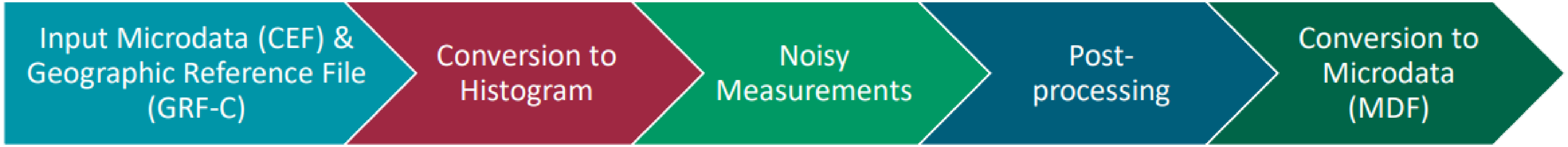
https://www.urban.org/sites/default/files/2022-09/Decennial%20Disclosure%20Explainer.pdf

Top Down Algorithm

Census Bureau’s implementation of a differential privacy framework



The Geographic Hierarchy


Input Microdata (CEF) & Geographic Reference File (GRF-C)
Conversion to Histogram
Noisy Measurements
Post-processing
Conversion to Microdata (MDF)

2020 DAS: Splitting person data from housing unit data

Person variables
Age
Sex
Race
Ethnicity
Relationship
Household/GQ type

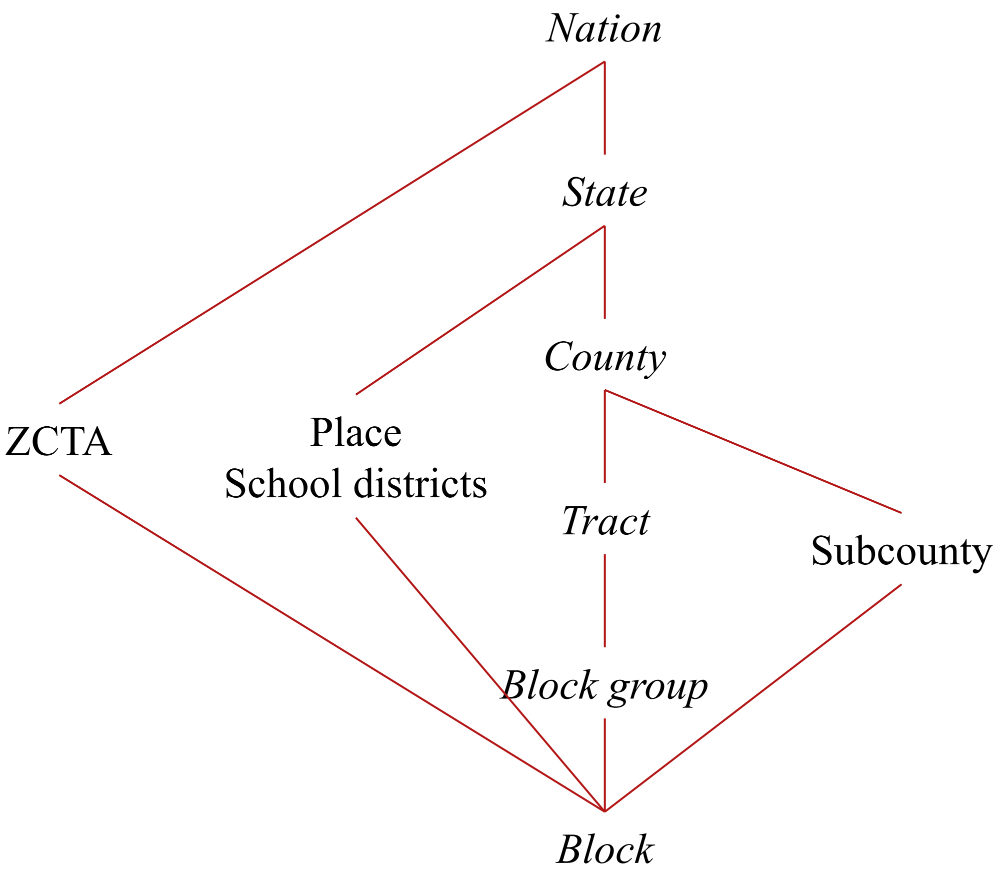
Housing unit variables
Occupancy status
Tenure
Race/ethnicity Hholder
Age HHolder
Household type
Household size

August 2022 demonstration data

- Application of latest TDA-settings to 2010 Census results
- Enables comparison with already released 2010 data
- Last round of feedback before final decisions are made
  - Formal feedback window has closed

Geographic "spine"

Traditional



Nation

State

County

ZCTA

Place
School districts

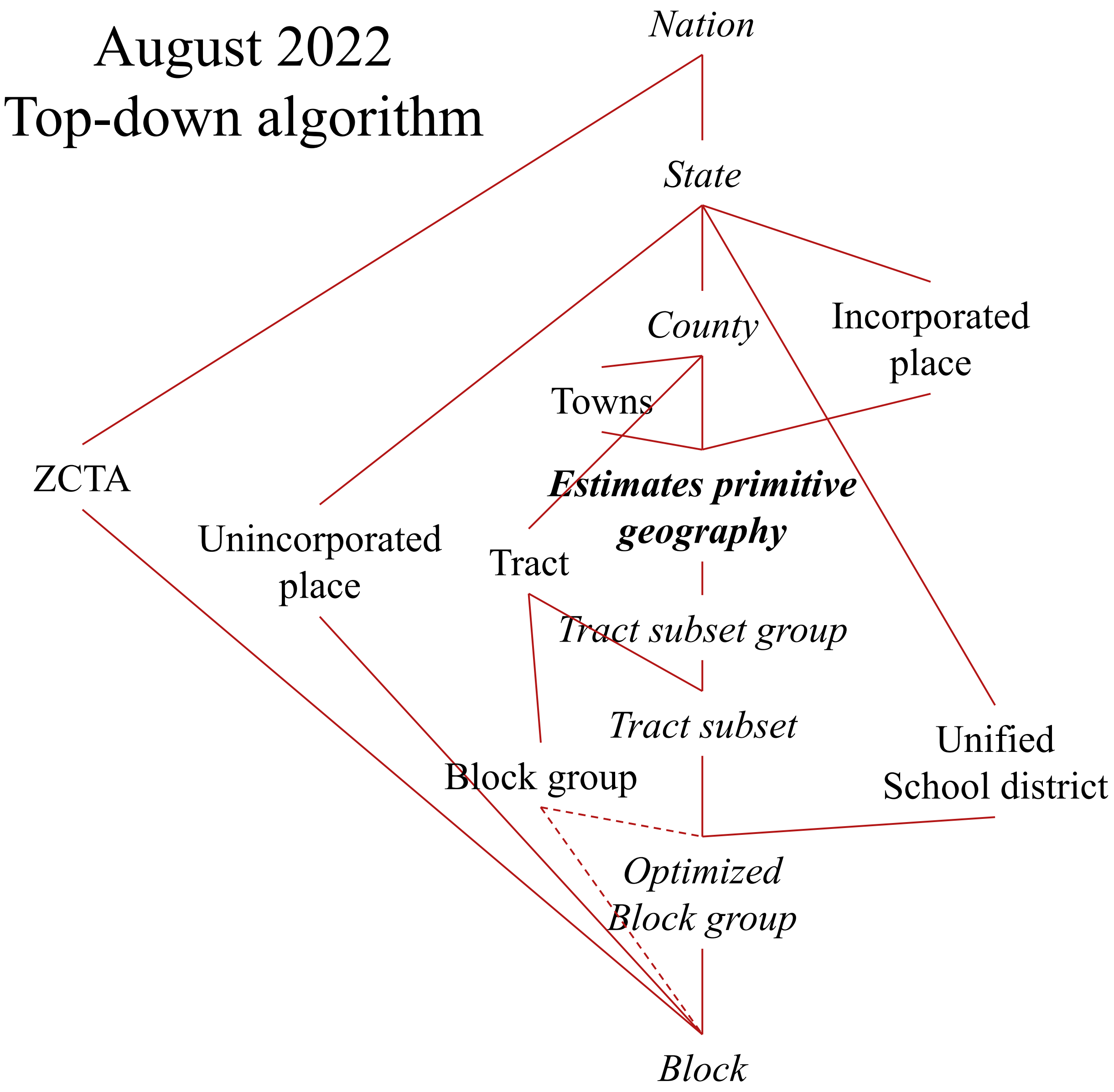
Tract

Subcounty

Block group

Block

August 2022
Top-down algorithm



Nation

State

County

Incorporated
place

Towns

ZCTA

Estimates primitive geography

Unincorporated
place

Tract

Tract subset group

Tract subset

Unified
School district

Block group

Optimized
Block group

Block

My block: age/sex composition

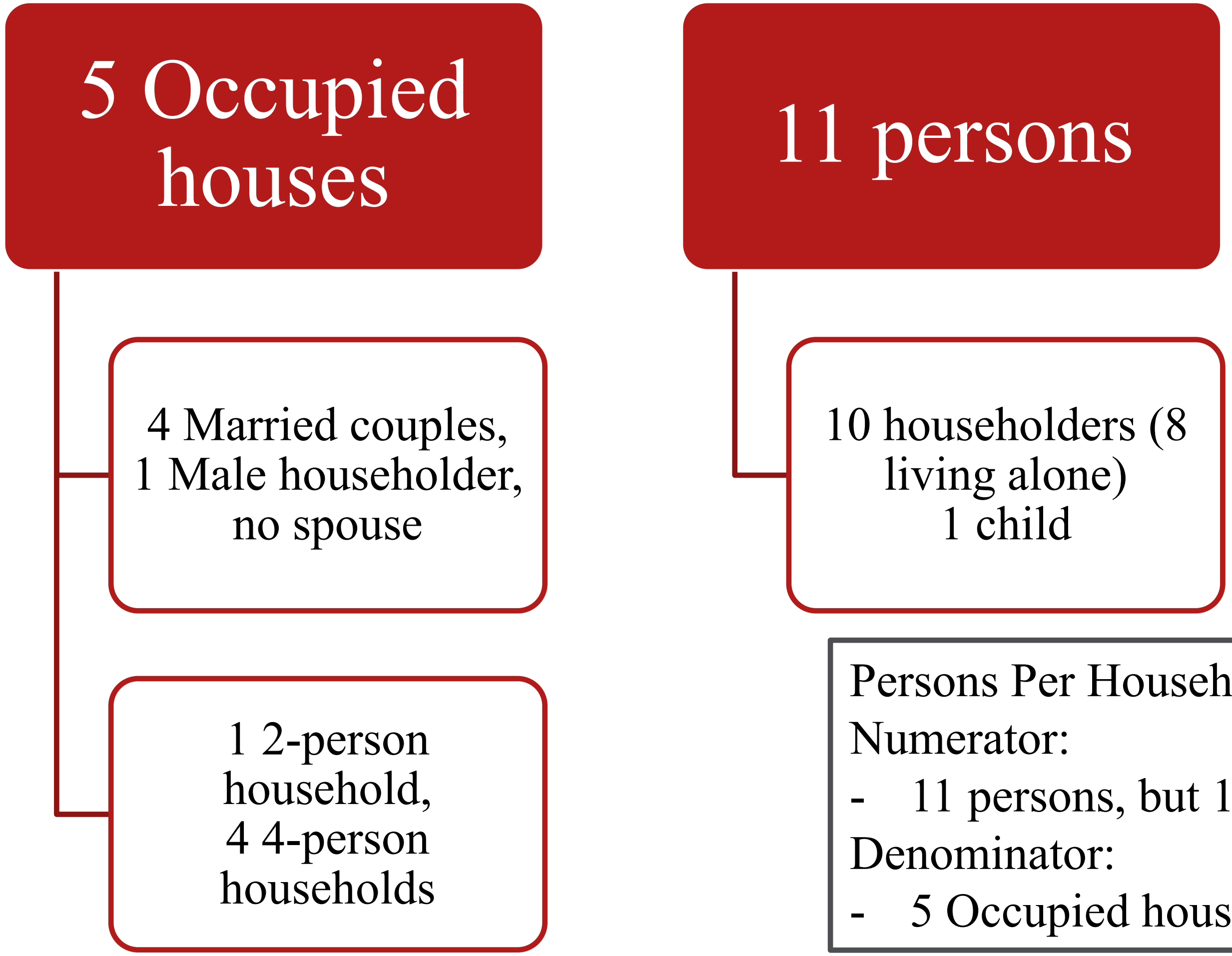
SF1 (all NonHisp White Alone)

	Males	Females
Age 10-14	1	
Age 18-19		1
Age 45-49	1	2
Age 50-54	1	
Age 80-84	1	1
Age 85+		1
Total	**4**	**5**

Demonstration data (August 22)

	Males	Females
Age 5-9	1 (NH Asian)	
Age 35-39	1 (NH Asian)	3 (NH White)
Age 40-44	1 (NH White)	
Age 65-66	1 (Hisp White)	
Age 80-84	3 (NH White)	
Age 85+	1 (NH White)	
Total	**8**	**3**

My Block inconsistencies



5 Occupied houses

4 Married couples,
1 Male householder,
no spouse

1 2-person household,
4 4-person households

11 persons

10 householders (8 living alone)
1 child

Persons Per Household?

Numerator:

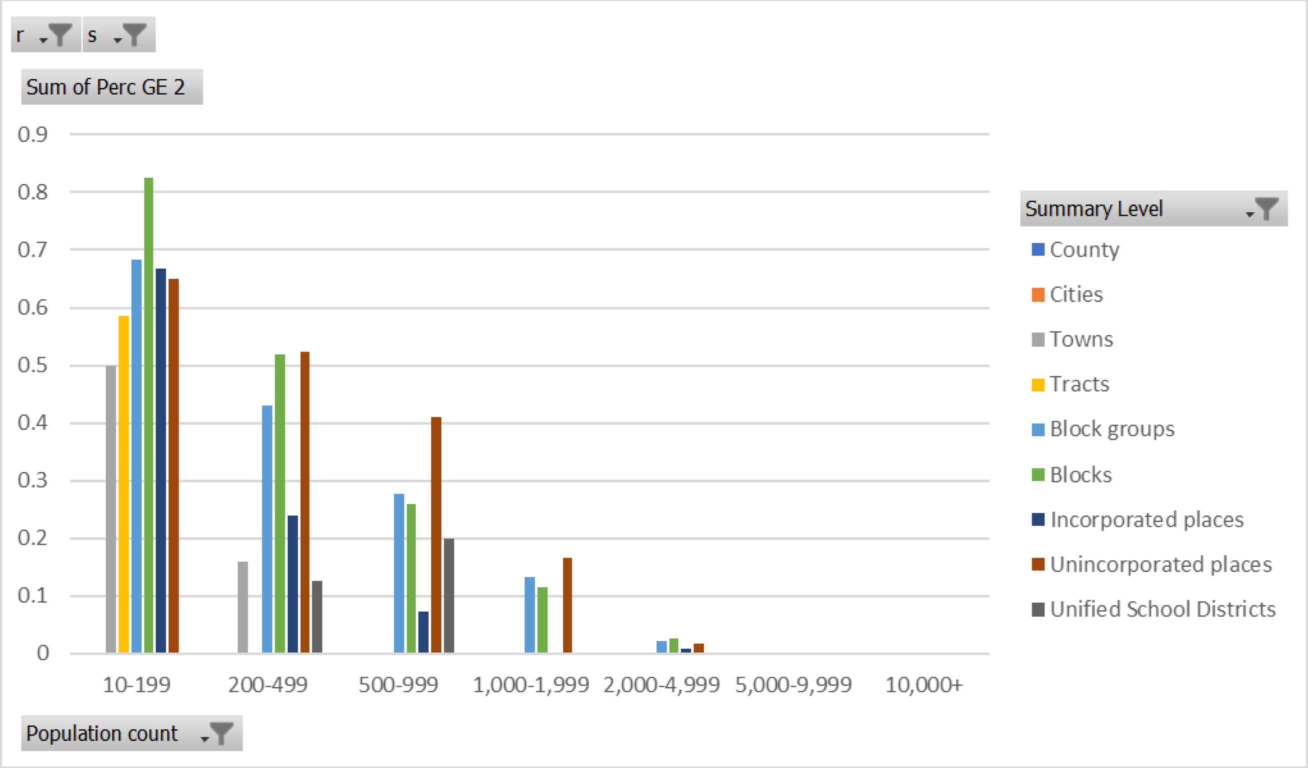
- 11 persons, but 18 in households by size

Denominator:

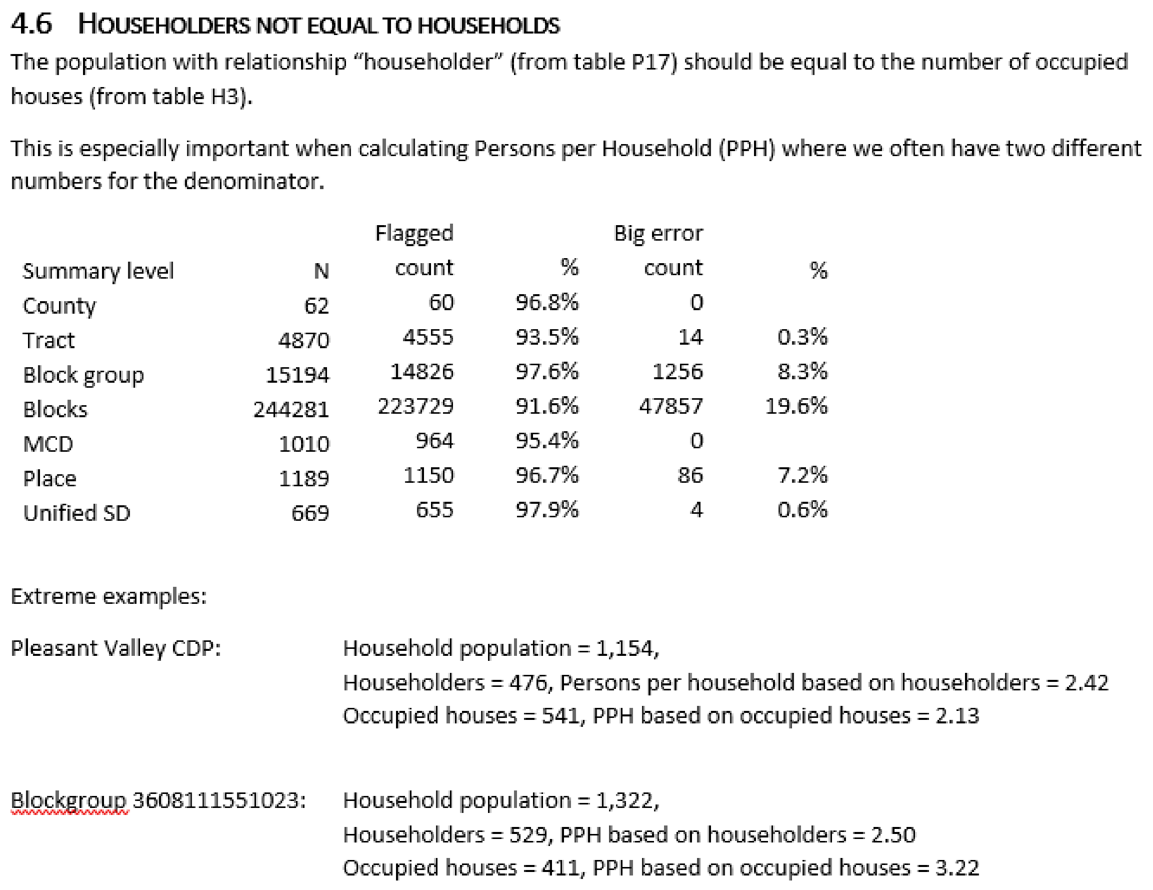
- 5 Occupied houses, but 10 householders

Findings latest version

Fraction of geographies with an absolute difference in median age of 2 yr or more



Inconsistencies



4.6 Householders not equal to households

The population with relationship "householder" (from table P17) should be equal to the number of occupied houses (from table H3).

This is especially important when calculating Persons per Household (PPH) where we often have two different numbers for the denominator.

Summary level	N	Flagged count	%	Big error count	%
County	62	60	96.8%	0	
Tract	4870	4555	93.5%	14	0.3%
Block group	15194	14826	97.6%	1256	8.3%
Blocks	244281	223729	91.6%	47857	19.6%
MCD	1010	964	95.4%	0	
Place	1189	1150	96.7%	86	7.2%
Unified SD	669	655	97.9%	4	0.6%

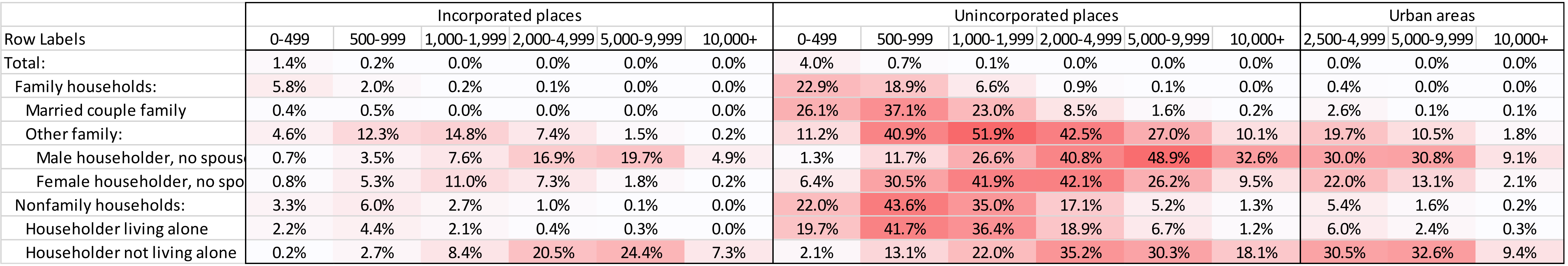
Extreme examples:

Pleasant Valley CDP:
Household population = 1,154,
Householders = 476, Persons per household based on householders = 2.42
Occupied houses = 541, PPH based on occupied houses = 2.13

Blockgroup 3608111551023:
Household population = 1,322,
Householders = 529, PPH based on householders = 2.50
Occupied houses = 411, PPH based on occupied houses = 3.22

Unincorporated places compared with incorporated places

Percent of geographies with ‘big errors’: difference >= 10, percent difference >= 10%



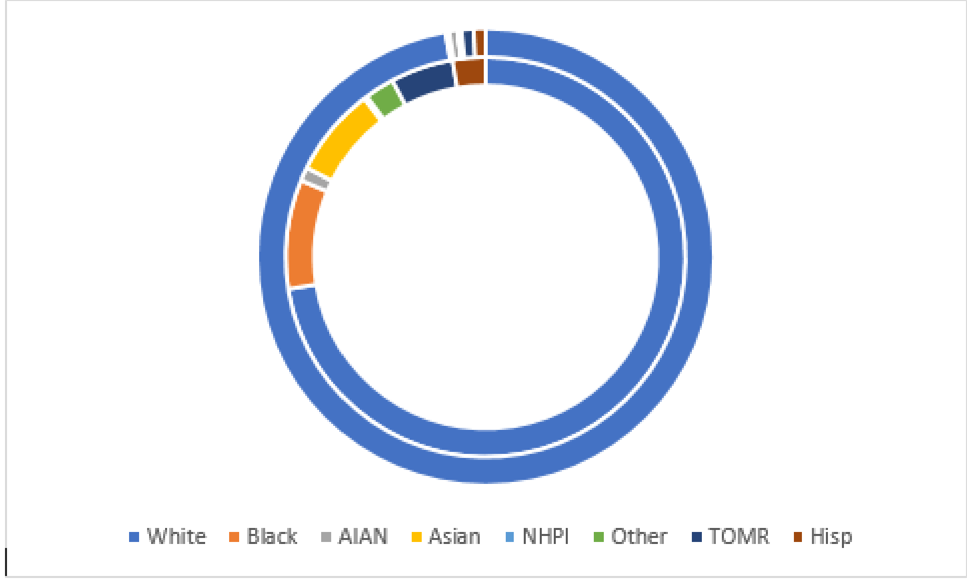
	Incorporated places						Unincorporated places						Urban areas		
Row Labels	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total:	1.4%	0.2%	0.0%	0.0%	0.0%	0.0%	4.0%	0.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Family households:	5.8%	2.0%	0.2%	0.1%	0.0%	0.0%	22.9%	18.9%	6.6%	0.9%	0.1%	0.0%	0.4%	0.0%	0.0%
Married couple family	0.4%	0.5%	0.0%	0.0%	0.0%	0.0%	26.1%	37.1%	23.0%	8.5%	1.6%	0.2%	2.6%	0.1%	0.1%
Other family:	4.6%	12.3%	14.8%	7.4%	1.5%	0.2%	11.2%	40.9%	51.9%	42.5%	27.0%	10.1%	19.7%	10.5%	1.8%
Male householder, no spous	0.7%	3.5%	7.6%	16.9%	19.7%	4.9%	1.3%	11.7%	26.6%	40.8%	48.9%	32.6%	30.0%	30.8%	9.1%
Female householder, no spo	0.8%	5.3%	11.0%	7.3%	1.8%	0.2%	6.4%	30.5%	41.9%	42.1%	26.2%	9.5%	22.0%	13.1%	2.1%
Nonfamily households:	3.3%	6.0%	2.7%	1.0%	0.1%	0.0%	22.0%	43.6%	35.0%	17.1%	5.2%	1.3%	5.4%	1.6%	0.2%
Householder living alone	2.2%	4.4%	2.1%	0.4%	0.3%	0.0%	19.7%	41.7%	36.4%	18.9%	6.7%	1.2%	6.0%	2.4%	0.3%
Householder not living alone	0.2%	2.7%	8.4%	20.5%	24.4%	7.3%	2.1%	13.1%	22.0%	35.2%	30.3%	18.1%	30.5%	32.6%	9.4%

Race householder of households with 3 generations

Example: Wyoming County, NY

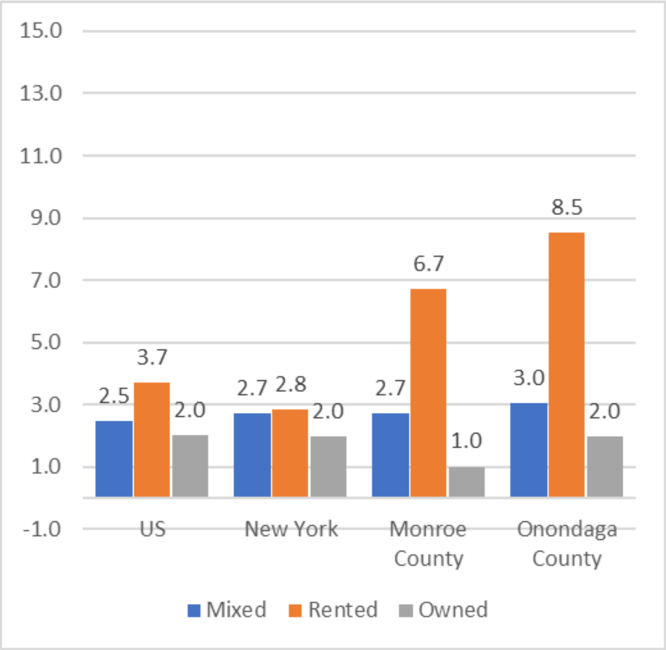
Outer ring SF1 race of householders of households with 3 or more generations (342 households). Inner ring DHC (271 households)

Figure 25: Race distribution of householders of households with three or more generations in Wyoming County, NY



Comparing accuracy by tenure

Households with Children (<18)



MdAPE of Census Tracts, by Aggregate Level

Lessons learned

- There are many inconsistencies, especially in smaller areas
- Never draw conclusions from a single block
- Geographies that are not brought closer to the spine have much more noise than those on, or closer to the spine
  - More population is needed for similar accuracy
  - Includes CDP, ZCTA, custom geographies
  - Rule of thumb: if ACS MOE's are generally large for your area of interest, assume that you need to be very careful interpreting DHC counts
- Try to avoid creating ratios, especially between person- and household variables

Guidance

- PRB handbook
- Shorter briefs in development for different audiences
- Final DHC demonstration dataset with final settings
  - Accuracy metrics based on this dataset

How will Differential Privacy affect data in the 2020 Census?

The Census Bureau has changed its privacy protection method to Differential Privacy. Instead of other methods formerly used, such as swapping, this method uses an algorithm called the Top-Down Algorithm (TDA) to randomly inject error into data, making it harder for computers to identify personal data. Almost all data in the redistricting product is subject to the TDA. Much of it can still be considered reliable, while other data points should be used with caution. Below is a guide produced by the Maine State Data Center for using data from the 2020 Redistricting Data product.

Green Light ●

This data is highly reliable. It has low levels of error and can be used as usual.

Green light variables:

- State-level total population, all other demographic & housing variables
- County-level total population, race & ethnicity except for Native Hawaiian and Pacific Islander, all other housing & Group Quarters variables
- Total population in medium and large cities & towns (>500 population)
- Total Population in Census Tracts
- Number and type of occupied group quarters units at the block level
- Total housing units at the block level
- Any data point at the county or county subdivision/place level with a count of at least 250
- Any data point at the Tract level with a count of at least 500

Yellow Light ●

This data is slightly less reliable, but can still be used with caution. Understand that these variables may have moderate error.

Yellow light variables:

- Race and ethnicity in large towns & cities (>5,000 population)
- Total population in small towns & cities
- Any cell at the county or county subdivision level that has a count between 60-250
- Any data point at the Tract level with a count between 250-500

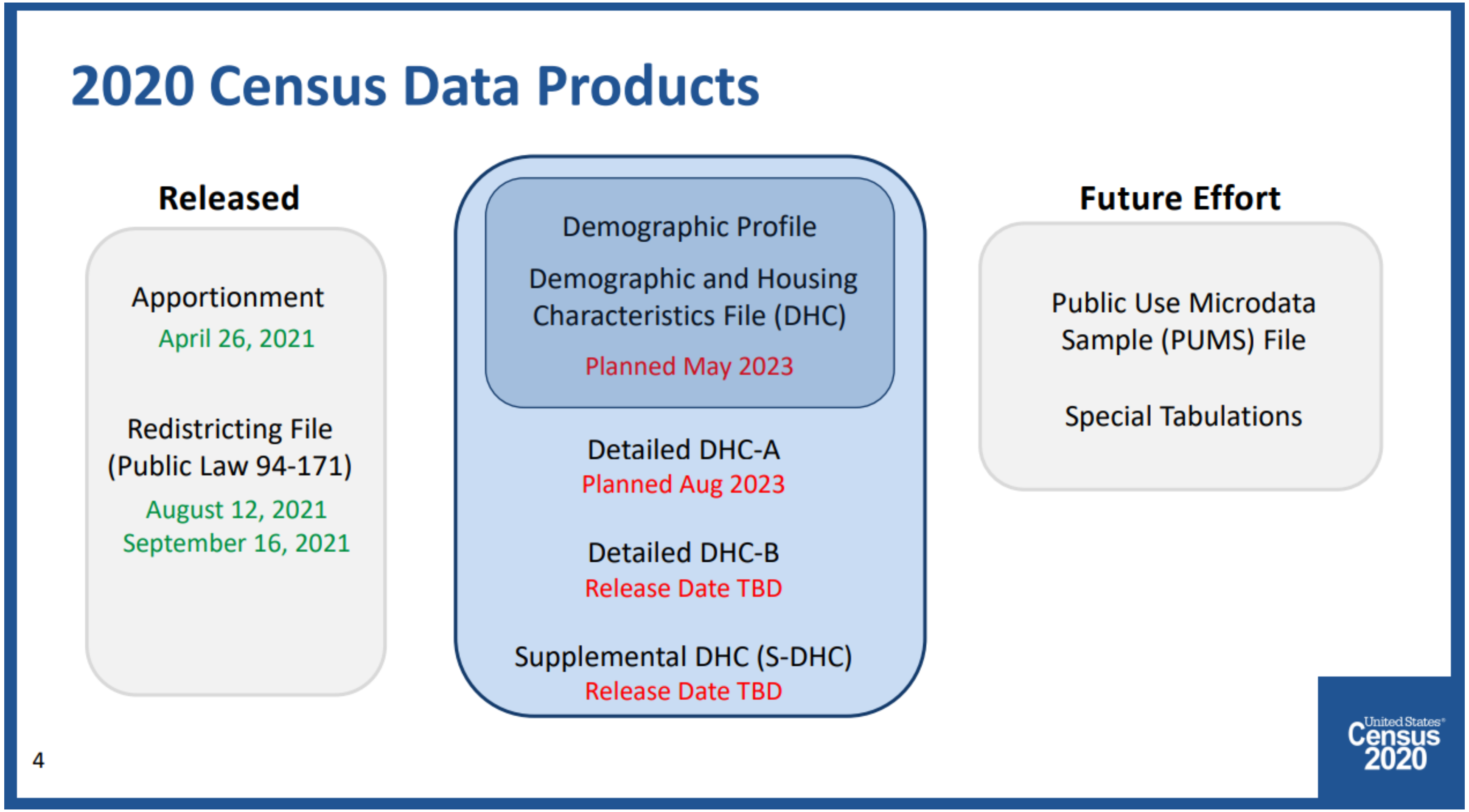
Red Light ●

This data is subject to high levels of error. Use these variables with extreme caution or consider aggregating them together to mitigate error.

Red light variables:

- Block-level data
- Race and ethnicity data in small towns
- Total population in county subdivisions with fewer than 60 population
- Any statistic that is divided across tables; for example, persons per household
- Any data point at the county subdivision level that has a count of less than 60
- Any data point at the tract level with a count of less than 250

Products schedule



2020 Census Data Products

Released

Apportionment
April 26, 2021

Redistricting File
(Public Law 94-171)
August 12, 2021
September 16, 2021

Demographic Profile

Demographic and Housing Characteristics File (DHC)

Planned May 2023

Detailed DHC-A
Planned Aug 2023

Detailed DHC-B
Release Date TBD

Supplemental DHC (S-DHC)
Release Date TBD

Future Effort

Public Use Microdata Sample (PUMS) File

Special Tabulations

4

Questions?