Differentially Private Algorithms for 2020 Census

Detailed DHC-A

Samuel Haney (**speaker)**, William Sexton, Ashwin Machanavajjhala

Michael Hay, Gerome Miklau

science@tmlt.io

TUMULT LABS

Disclaimer

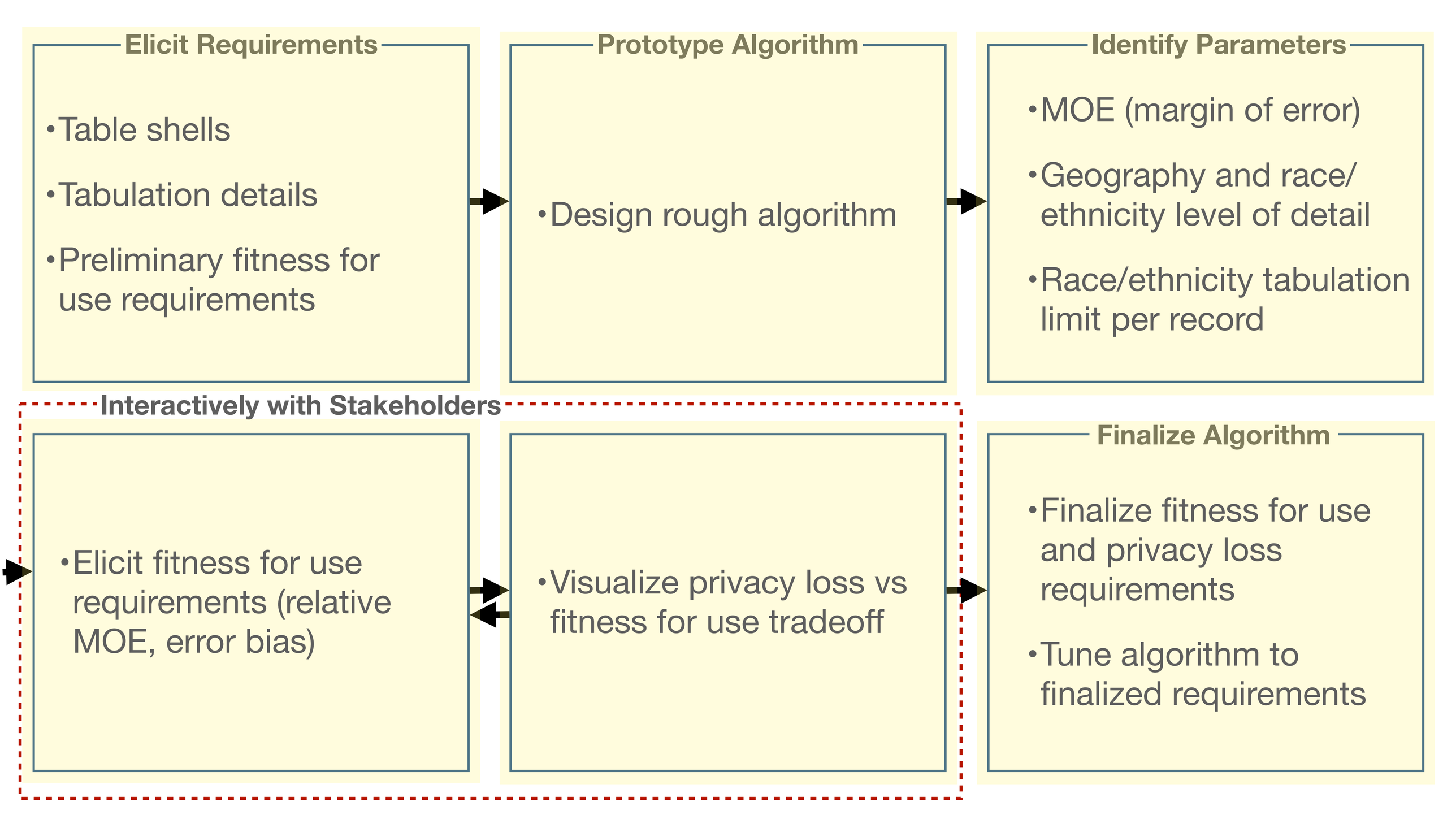
- Tumult has been contracted to develop a candidate algorithm for the detailed race and ethnicity product for the 2020 decennial census, but the Census Bureau has not yet decided to use this algorithm.

- In order to compare the two algorithms based on Geometric and Discrete Gaussian mechanisms, we consider candidate accuracy targets (and the corresponding privacy loss budgets), some of which were setup in consultation with US Census Bureau SMEs. No determination has been made by the US Census Bureau about either accuracy targets or privacy loss budgets for this data product.

Takeaways

- Privacy-utility negotiation takes time and effort

- Specific tools and techniques we use


Elicit Requirements
•Table shells
•Tabulation details
•Preliminary fitness for use requirements
Prototype Algorithm
•Design rough algorithm
Identify Parameters
•MOE (margin of error)
•Geography and race/ethnicity level of detail
•Race/ethnicity tabulation limit per record
Interactively with Stakeholders
•Elicit fitness for use requirements (relative MOE, error bias)
•Visualize privacy loss vs fitness for use tradeoff
Finalize Algorithm
•Finalize fitness for use and privacy loss requirements
•Tune algorithm to finalized requirements

Takeaways

- Privacy-utility negotiation takes time and effort

- Specific tools and techniques we use

record:

geo_id	sex	age	race1	race2	…	race8	eth
10000	F	81	1001	5934		Null	1007

Determines mapping to ≤ 4 geographies

Determines mapping to ≤ 18 detailed races/ethnicities

North Carolina,
Egyptian Alone

Total:	0
F, 0-4:	0
F, 5-9:	0
…	…
F, 80-84:	0
F, 85+:	0
M, 0-4:	0
…	…

Orange County,
Japanese Alone or in Combo

Total:	0
F, 0-4:	0
F, 5-9:	0
…	…
F, 80-84:	0
F, 85+:	0
M, 0-4:	0
…	…

California,
Aleut Alone or in Combo

Total:	0
F, 0-4:	0
F, 5-9:	0
…	…
F, 80-84:	0
F, 85+:	0
M, 0-4:	0
…	…

One table per population group = Geography x Detailed Race/Ethnicity

TUMULT LABS

Additional criteria

- Relative error is main utility measure
- We care about intermediate breakouts (sex marginal)
- Statistics must be integral, but there are no other consistency requirements

Simple Algorithm

- Add discrete Laplace noise to each statistic, sensitivity is (max geographies per record) x (max race/ethnicity groups per record) x (# statistics each record contributes to) = 144.

- We next present a series of modifications to this basic algorithm.

Optimization 1

Adjust the privacy budget separately for separate race/ethnicity and geography *levels*.

Nation: USA

State: California,
North Carolina,
…

County: Orange County,
Durham County,
…

AIANNH: Allegany Reservation
…

Regional: European,
North African,
…

Detailed: Albanian,
Egyptian,
…

Larger groups on average = less budget
More important groups = more budget

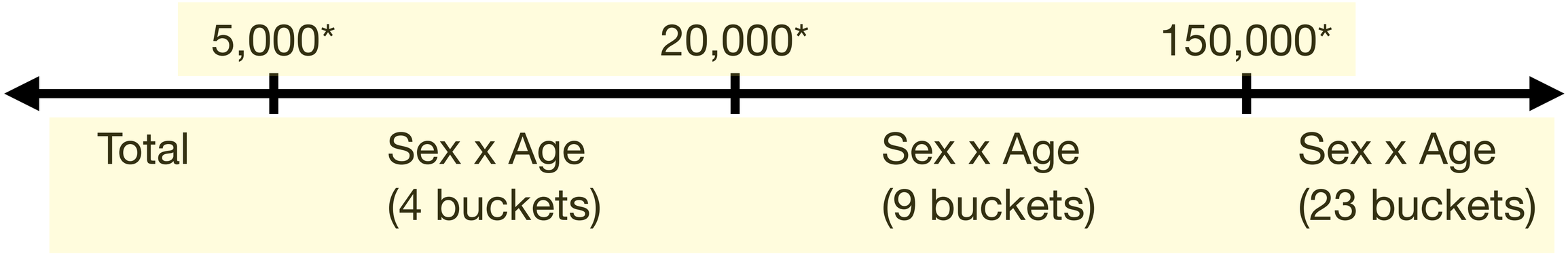
Smaller groups on average = more budget

Maximum contribution of one person within each geo level x race/ethnicity level is 9.

TUMULT LABS

Optimization 2

Adaptively choose the statistics level, using a fraction of the privacy budget.



(European, California)

Size: 21,000

Noisy Size: 19,000

TUMULT LABS

*These are not the thresholds that will be used in production.

Optimization 3

Use discrete Gaussian mechanism and zCDP privacy accounting.

- Alternate privacy definition that can be converted to approximate differential privacy.
- Performs well compared to differential privacy when composing many queries.

TUMULT LABS

*These are not the privacy loss values that will be used in production.

Optimization 3

Use discrete Gaussian mechanism and zCDP privacy accounting.

$\varepsilon = 15.3$

pure differential privacy

$\varepsilon = 12.2$

approximate differential privacy with $\delta = 10^{-10}$

TUMULT LABS

*These are not the privacy loss values that will be used in production.

Thank you!

Questions?