Per-Record Differential Privacy and the Census of Agriculture

Michael Jacobsen, William Sexton, Casey Meehan, Ashwin Machanavajjhala, Yang Cheng

October 24, 2024

2024 FCSM Research and Policy Conference

Disclaimer

- The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Motivation

- NASS is investigating new statistical disclosure methods and their possible impacts on the Census of Agriculture
- Complementary cell suppression (Cox, 1995) is the current methodology
  - Primary selection uses p-percent rule
  - Prevents the other records from learning about specific value of primary suppression
  - May lead to oversuppression and lack of utility through:
    - Too many suppressed cells
    - Too much data suppressed
  - Privacy parameters are not published → lack of user transparency

2017 Census of Agriculture

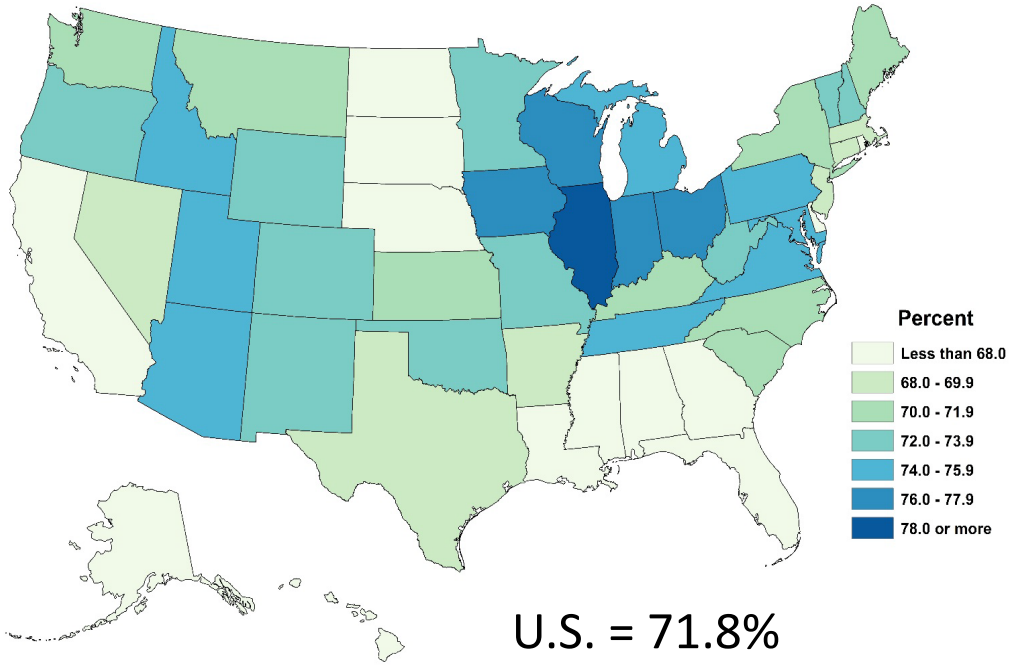
Census Mailing List

~3M records

Not on Mail List

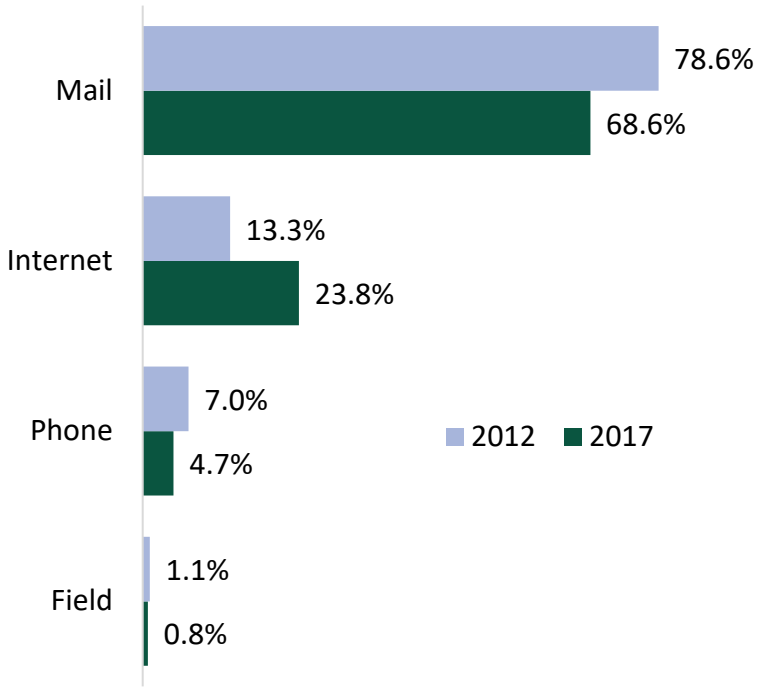
~42K records

Response Rate by State, 2017



U.S. = 71.8%

Return Rate by Mode, 2012 and 2017 *(percent of returns)*



Collected data edited, weighted and summarized prior to disclosure

Case Study

- Disclosure avoidance applied to simulated dataset that resembles the 2017 Michigan Chapter 2, Table 31 (Fruits and Nuts)
- This table embodies key privacy challenges
  1) Small county-level sums are hard to protect (64% suppression under current suppression method)
  2) High skewness - some cells dominated by a few farms

Table 31. **Fruits and Nuts: 2017 and 2012** (continued)

[For meaning of abbreviations and symbols, see introductory text.]

Geographic area		Total		Bearing age acres		Nonbearing age acres	
		Farms	Acres	Farms	Acres	Farms	Acres
PEARS, BARTLETT							
State Total							
Michigan	2017	242	459	153	413	110	47
	2012	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Counties, 2017							
Alcona		1	(D)	1	(D)	-	-
Allegan		11	18	9	16	4	2
Alpena		3	(D)	1	(D)	2	(D)
Antrim		1	(D)	1	(D)	-	-
Barry		2	(D)	2	(D)	-	-
Bay		4	1	-	-	4	1
Benzie		4	(D)	2	(D)	4	(D)
Berrien		14	36	14	36	-	-
Branch		1	(D)	1	(D)	-	-
Calhoun		7	3	7	2	3	1

Differential Privacy (DP)

- **Why DP? -** Provides quantifiable privacy protection against strong adversarial models.
- DP does not require suppression and allows for transparency.

DP Deployments

Census Bureau

IRS

Wikimedia

Growing recognition of DP in the Federal Statistical System

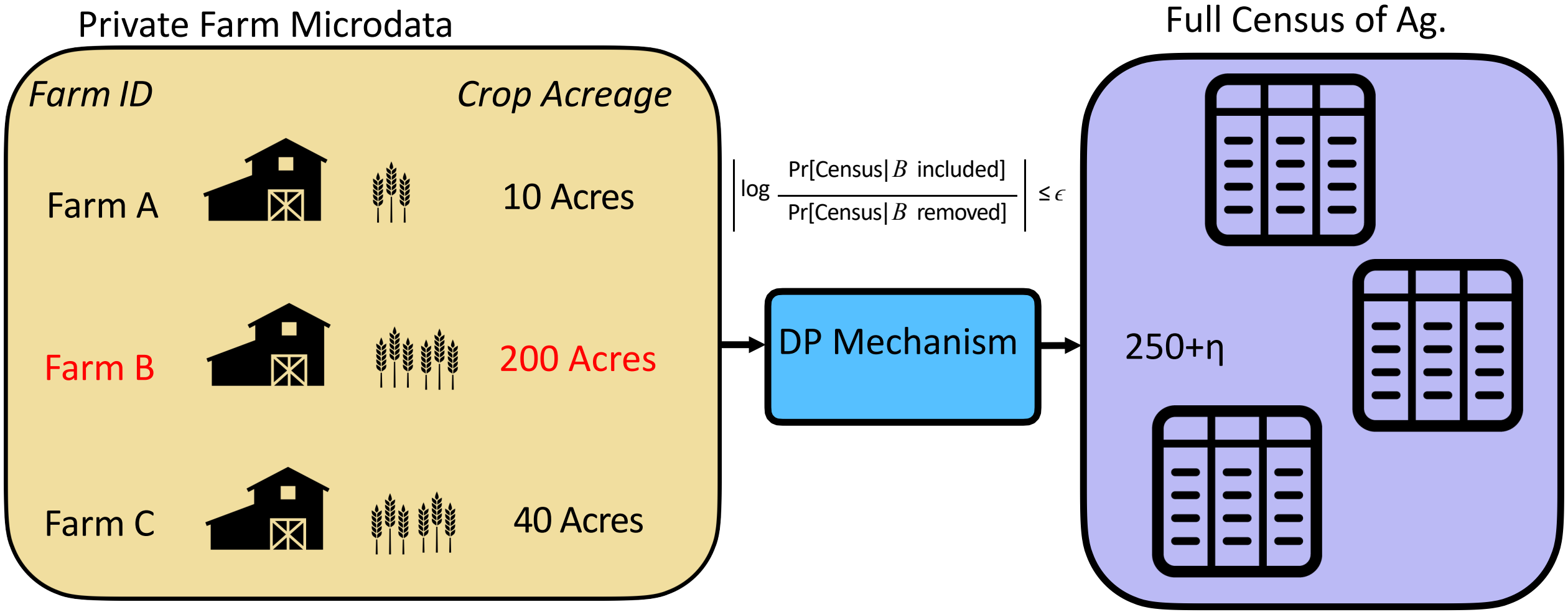
Executive Order 14110 (Oct 30, 2023): Sec 9(b)

"[…] NIST shall create guidelines for agencies to evaluate the efficacy of **differential-privacy-guarantee** protections, including for AI […]"

FCSM Conference

Sessions, workshops, presentations devoted to DP

Differential Privacy



Differential Privacy

- DP has poor privacy/utility tradeoff on highly skewed data
- Strong privacy and acceptable utility often not possible when cell is dominated by a few records
- These issues are exacerbated by weighted data

Farm Microdata

Farm ID	Crop Acreage
A	10
B	200
C	40

Low privacy loss

$\varepsilon = 1$

High privacy loss

$\varepsilon >> 10$

County-Level
Census Aggregate Data

Crop Acreage

$250+\eta = 1{,}051$ acres

$250+\eta = 267$ acres

Per-Record Differential Privacy

***Per-Record Differential Privacy (PRDP)* is a generalization of standard DP**

PRDP was developed to offer nuanced privacy guarantees to highly-skewed data.

PRDP is an emerging formal privacy notion

- Provides quantifiable privacy protection against strong adversarial models.
- Does not require suppression and allows for increased transparency.
- Provides sliding protection that enables better utility on skewed data.
- Captures privacy impact of weighted data.

PRDP Methodology

1. Test different privacy-loss budgets ε=1 and ε=2
2. Set the privacy threshold parameter T_a
   - x_a = weighted record acreage value for commodity *a*
   - T_a = median x_a for records with *a* > 0
   - Farms with $x_a < T_a$ receive ε privacy loss
   - Farms with $x_a > T_a$ receive $(x_a / T_a) * \varepsilon$ privacy loss
3. Add Laplace noise η_c with scale T_a / ε to cell *c*'s true value v_c
4. (Optional) Suppress overly noisy data
   - Suppress cell *c* if noisy value $v_c + \eta_c \leq k * \sigma_c$, where
     - σ_c = std. deviation of the Laplace noise distribution
     - k>0

Formal privacy guarantee, better utility?

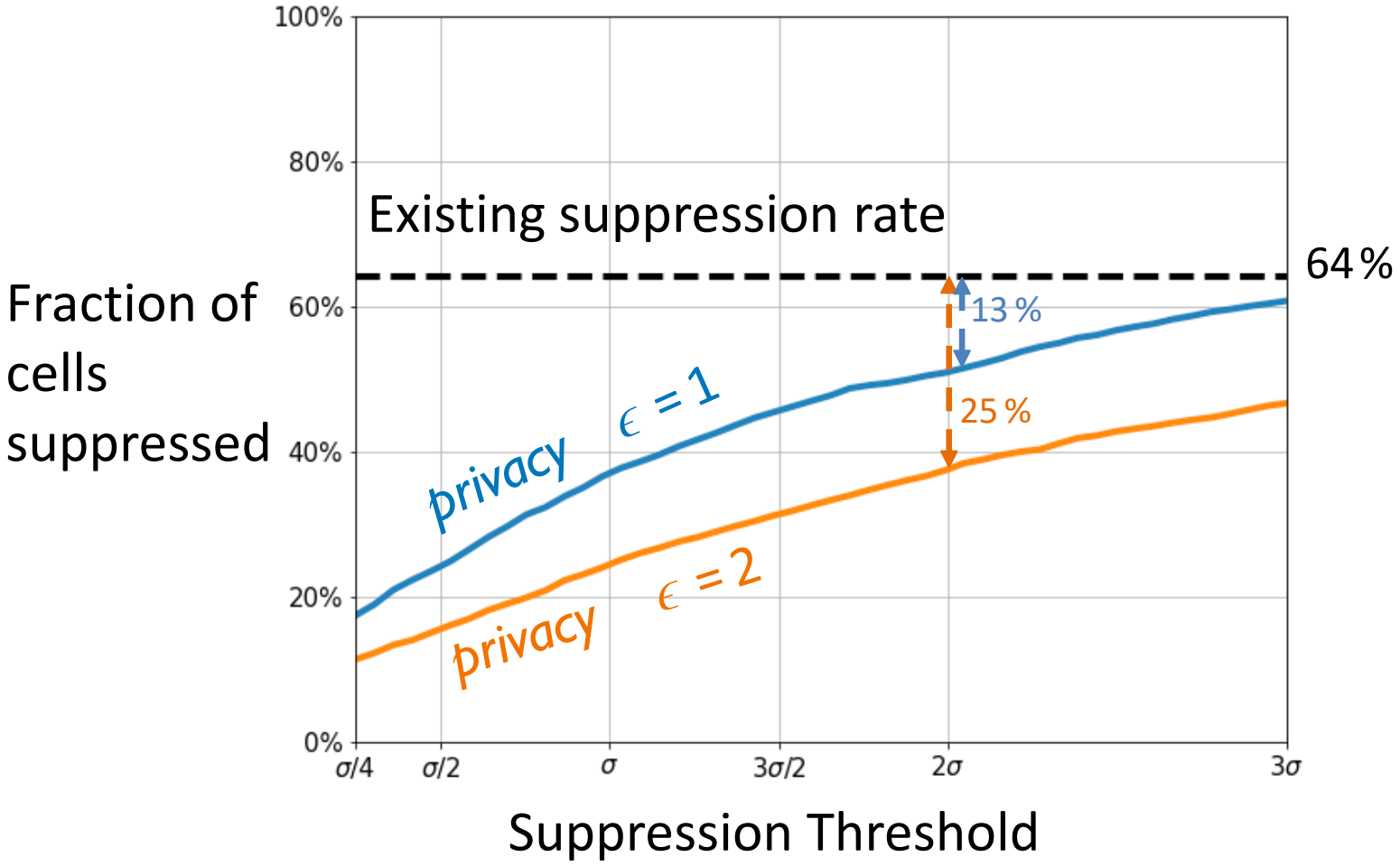
Primary Questions

1) **Can we release more cells?**
   Too many cells currently suppressed (64% suppressed in Table 31)
   Can we release more cells to data users with PRDP?

2) **Utility of (noisy) released cells?**
   Unsuppressed cells have added noise.
   Are these cells still accurate/useful?

PRDP Impact on Suppressing Cells

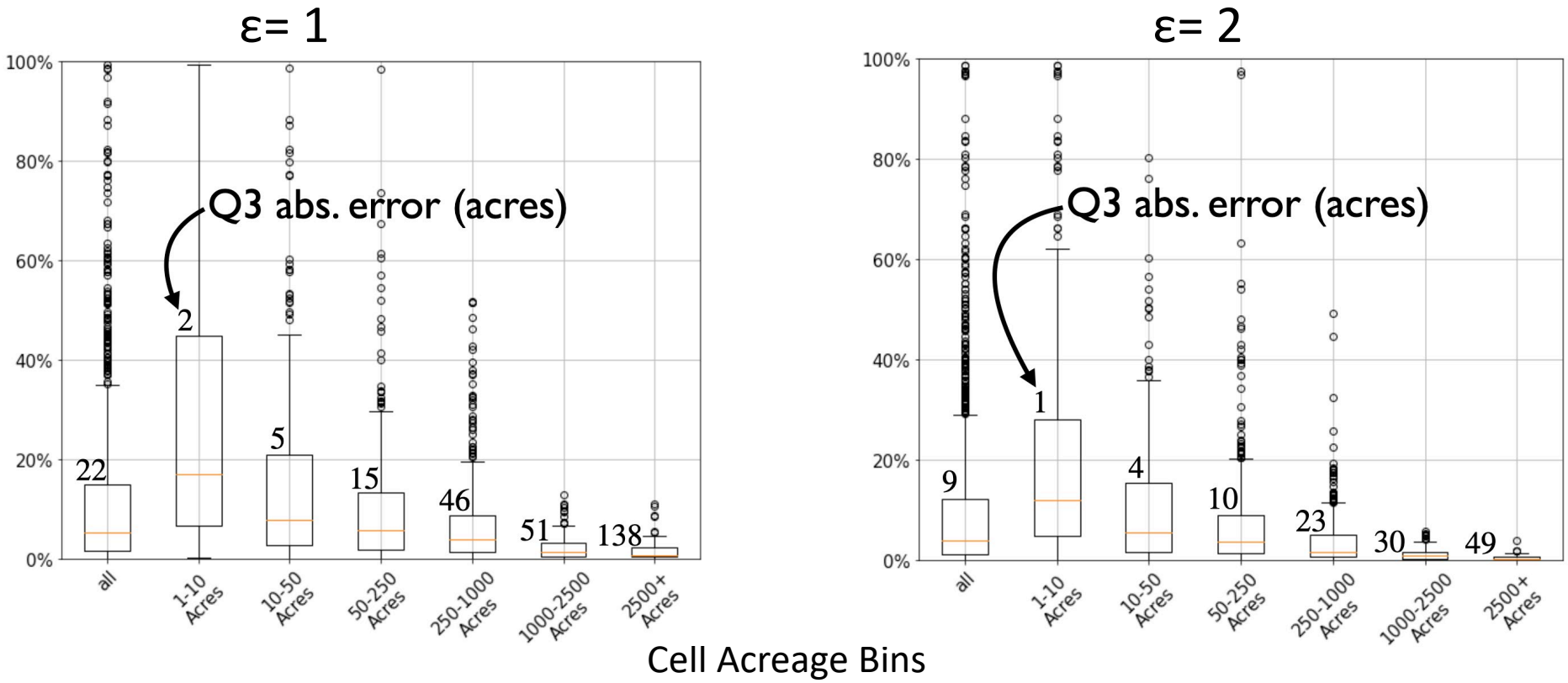
- Number of cells suppressed decreases under PRDP, with a larger decrease coming with increased ε



PRDP and Relative Error

- Low values: cell size ranges are maintained
- High values: cell values are maintained

Relative Noise Error
$=| \eta_c |/ v_c$



In Conclusion

- Differential Privacy is a forward-looking disclosure avoidance approach
  - Better than cell suppression for privacy, utility, and transparency
  - Growing acceptance in the federal statistical system
- PRDP adapts DP-style guarantees to Census of Ag's highly skewed data
- Case study on Michigan Table 31 simulated data
  - Improved suppression rate from 64 % → 40 %
  - Evidence of low noise for unsuppressed cells – further evaluation

References

Cox, Lawrence H. "Network models for complementary cell suppression." *Journal of the American Statistical Association* 90.432 (1995): 1453-1462.

Dwork, Cynthia. "Differential privacy." *International colloquium on automata, languages, and programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

Executive Order 14110 Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *Federal Register* https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and- trustworthy-development-and-use-of-artificial-intelligence (2023).

Kelly, James P., Bruce L. Golden, and Arjang A. Assad. "Cell suppression: Disclosure protection for sensitive tabular data." *Networks* 22.4 (1992): 397-417.

Seeman, Jeremy, et al. "Privately Answering Queries on Skewed Data via Per Record Differential Privacy." *arXiv preprint arXiv:2310.12827* (2023).