Differentially Private Methods for Counts

SDSS 2022

Statistical Data Privacy Techniques for Sharing Sensitive Data

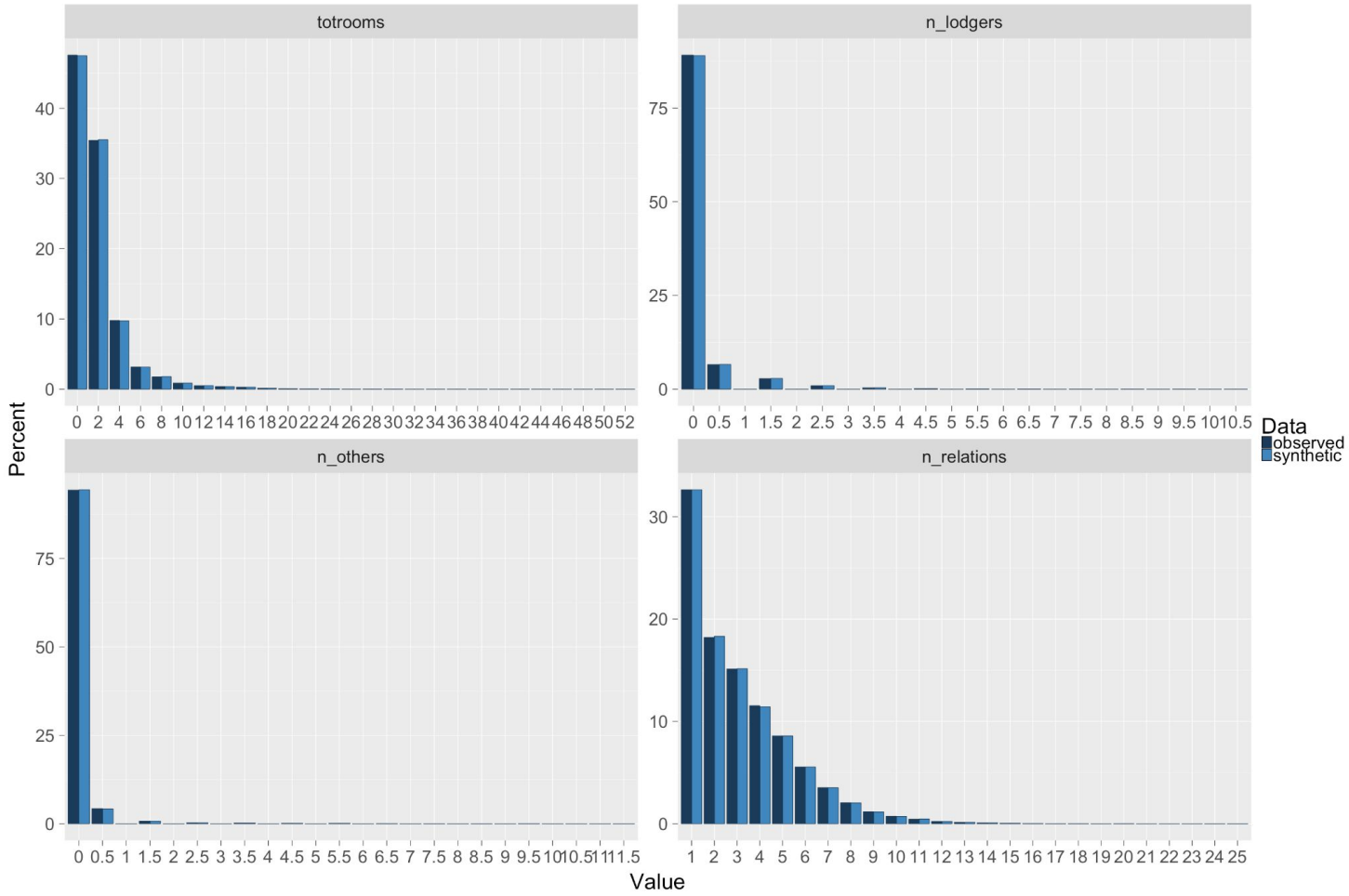
Short Course: Part 4

Differential Privacy is an Entirely New Framework

Motivation for a New Privacy Framework

- Recall from previous lessons:
  - Assumptions on population data
  - Assumptions on attacker knowledge on participation
  - Assumptions on publicly accessible variables
  - Lack of consensus on risk definitions for synthetic data

Values are Synthetic but are They Protective?



Goal is Meet a Provable Definition of Privacy

- DP is not a method, it is a set of *definitions*, e.g.:

A randomized algorithm, $\mathcal{M}$, is ϵ-Differentially Private if for all $S \subseteq Range(\mathcal{M})$ and for all X, X' such that $\delta(X, X') = 1$:

$$\frac{P(\mathcal{M}(X') \in S)}{P(\mathcal{M}(X) \in S)} \leqslant exp(\epsilon).$$

- **ε**-DP in words:
  - The relative likelihood of publishing any given statistic value is bounded given the presence or absence of any individual
- Does not require assumptions about the attacker
- Must understand space of all possible observations

Let's Try Out an Example

- Suppose we want to release a count of COVID cases

County	Sex	Race/Ethnicity	Case Count w/ person X	Case Count w/o person X
Allegheny	F	W	101	100
Allegheny	M	W	683	683
…	…	…	…	…

- With no noise added, the ratio of likelihoods is unbounded
  - W/ person X $\rightarrow$ release count = 101 w.p. 1, release count = 100 w.p. 0
  - W/o person X $\rightarrow$ release count = 100 w.p. 1, release count = 101 w.p. 0

Let's Try Out an Example

- Add random noise to the counts to bound the relative likelihood

County	Sex	Race/Ethnicity	Case Count w/ person X	Case Count w/o person X
Allegheny	F	W	101 + Lap(Δ/ε)	100 + Lap(Δ/ε)
Allegheny	M	W	683	683
…	…	…	…	…

Let's Try Out an Example

- Add random noise to the counts to bound the relative likelihood

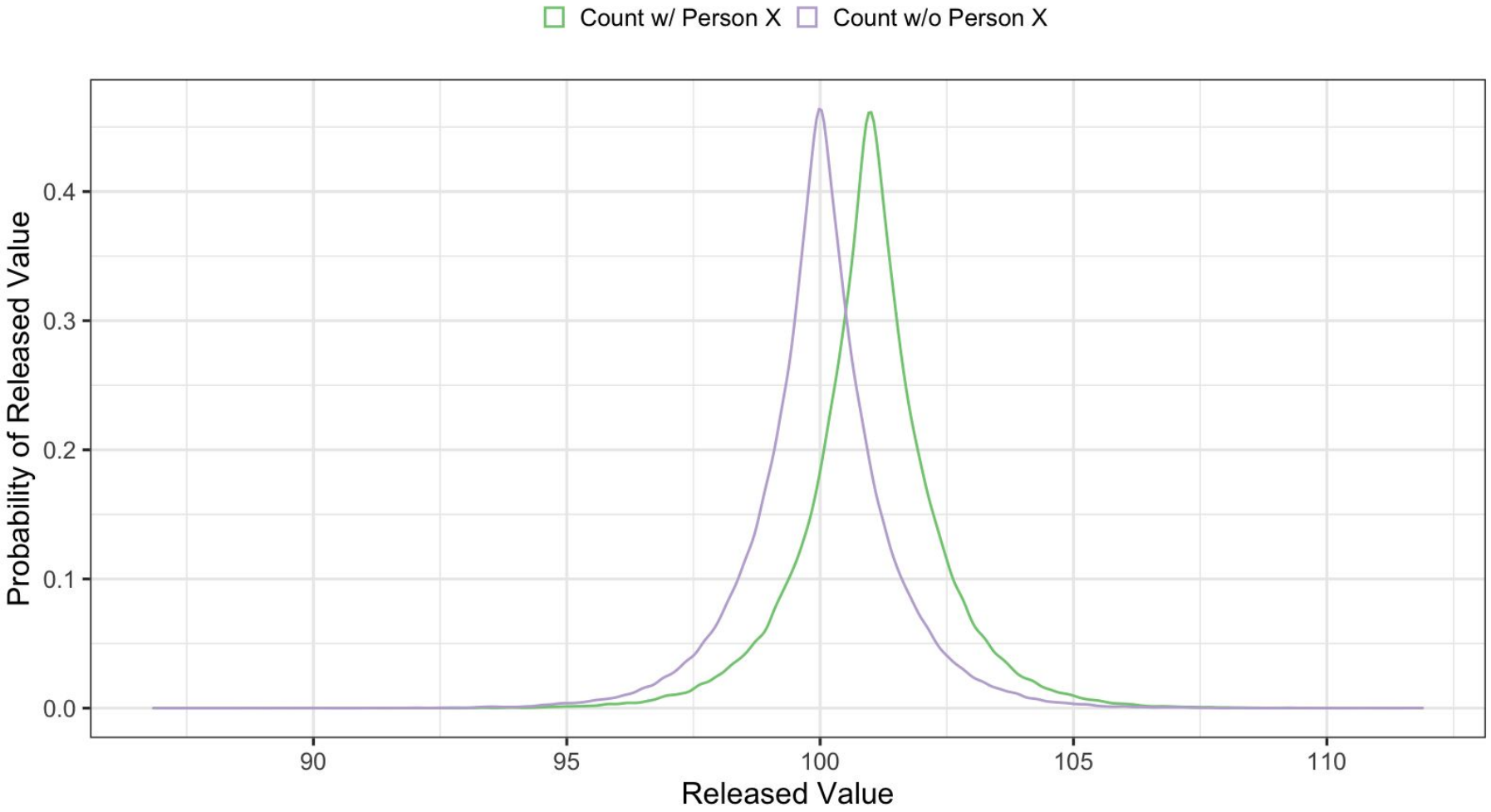
County	Sex	Race/Ethnicity	Case Count w/ person X	Case Count w/o person X
Allegheny	F	W	101 + Lap(Δ/ε)	100 + Lap(Δ/ε)
Allegheny	M	W	683	683
…	…	…	…	…

- Proven that Laplace noise meets the definition of ε-DP
  - Laplace scale parameter (variance) must: Δ / ε

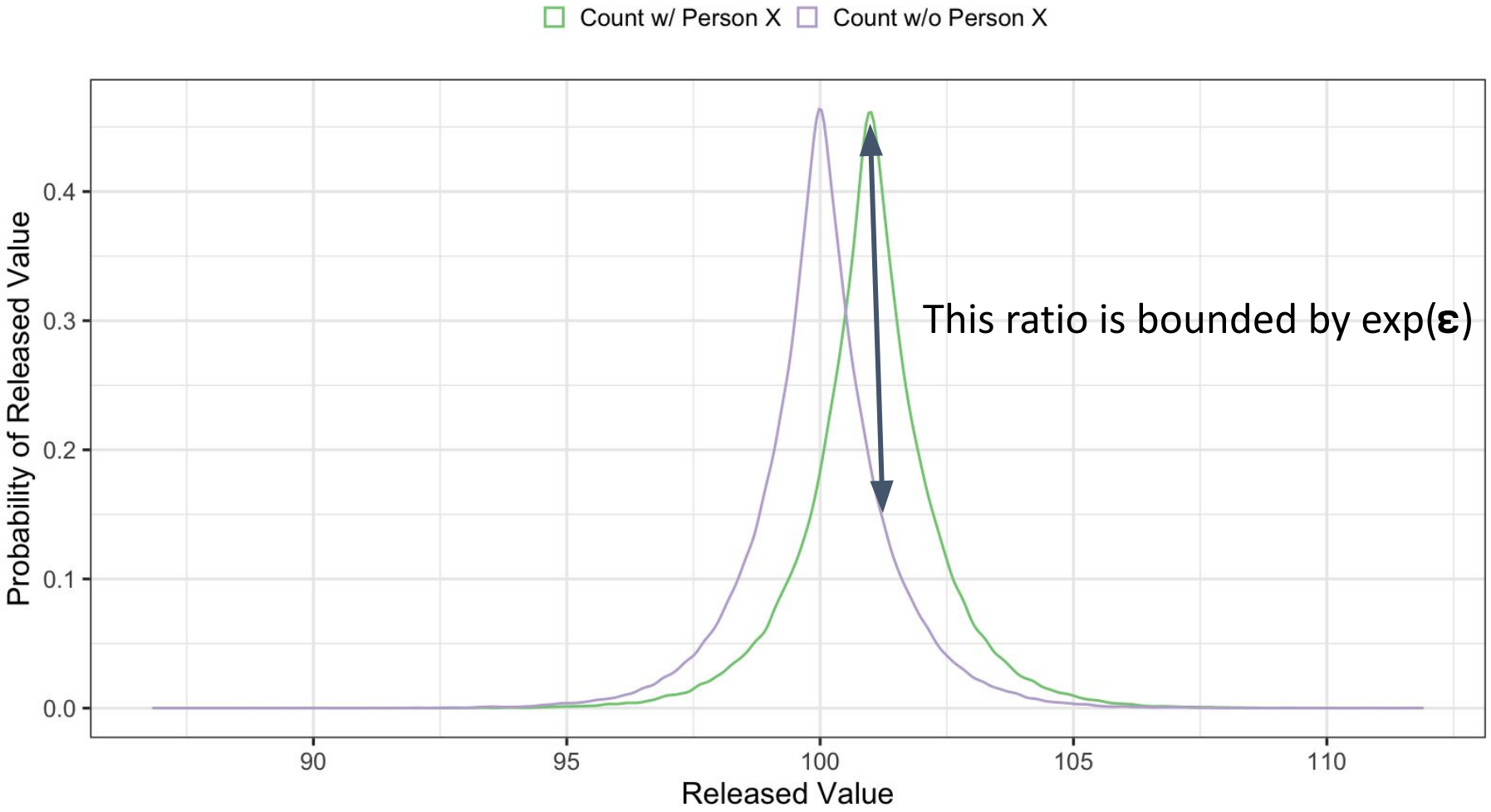
Max possible difference in count due to one person.
In this case Δ = 1 (Global sensitivity)

Privacy parameter

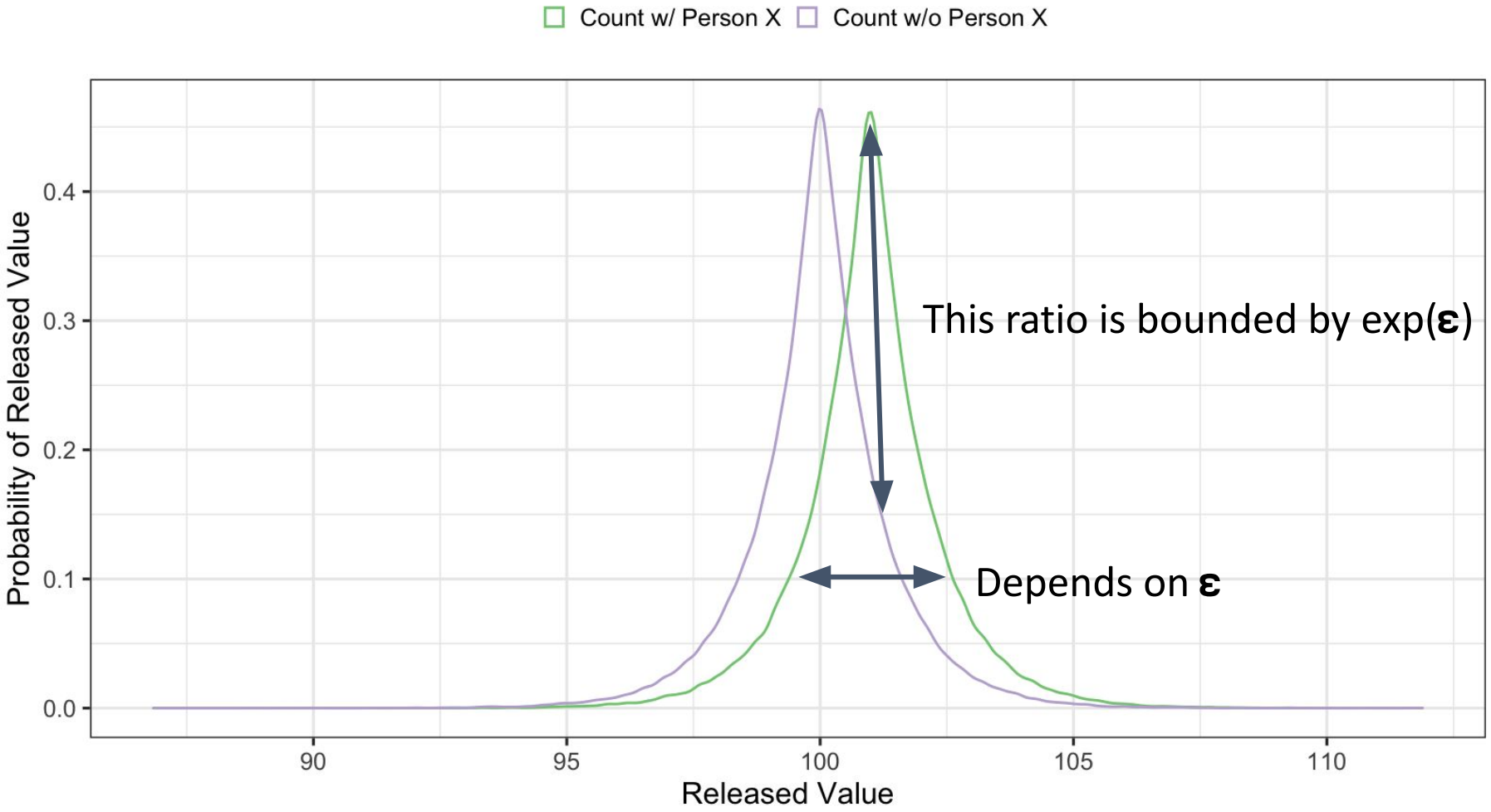
We are Bounding the Difference Between Two Distributions of Possible Outputs



We are Bounding the Difference Between Two Distributions of Possible Outputs



The Variance Depends on the Privacy Loss (and the Sensitivity)



We Can Release Multiple Counts

- If counts are from disjoint subsets, no need to add **ε**

County	Sex	Race/Ethnicity	Case Count
Allegheny	F	W	101 + Lap(1/**ε**)
Allegheny	M	W	683 + Lap(1/**ε**)

- If counts are based on the same individuals, add **ε**

County	Sex	Race/Ethnicity	Case Count	Death Count
Allegheny	F	W	101 + Lap(1/ε_1)	1 + Lap(1/ε_2)

Total $\varepsilon = \varepsilon_1 + \varepsilon_2$

- Leads to the idea of a “privacy budget”
- More complex composition theorems exist

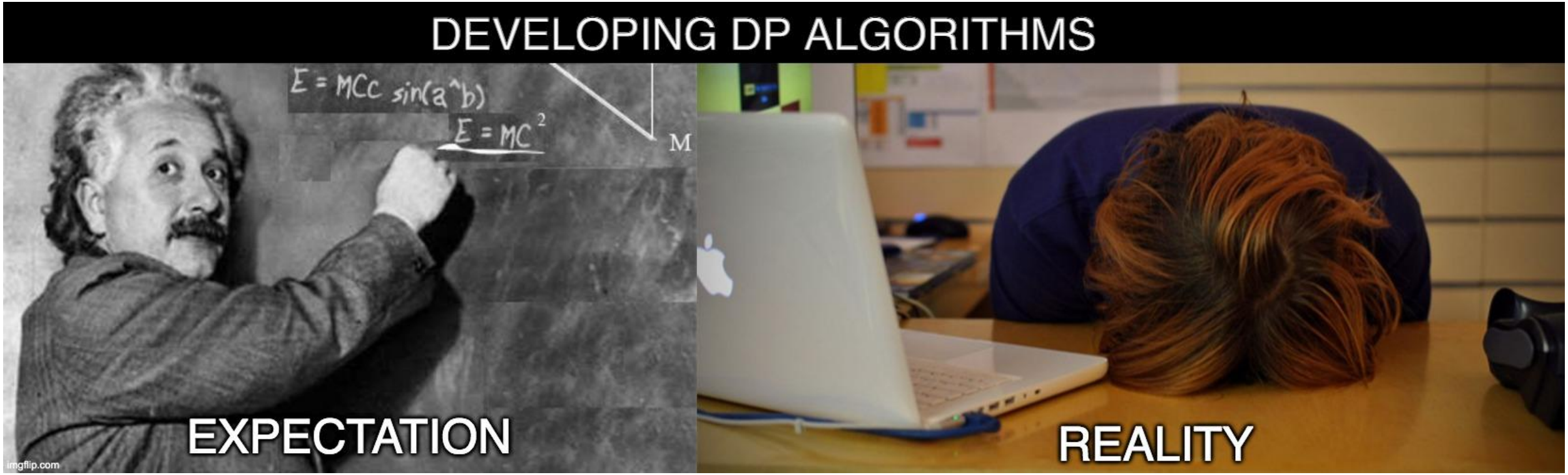
Noisy Results can be Post-Processed

- Any function of an output that satisfies **ε**-DP is also **ε**-DP
- Post-processing means applying other functions to the data
  - Rounding the counts
  - Enforcing no counts to be negative
  - Rescaling so that total counts remain fixed (invariants)
- **Important: Cannot use information from the confidential data**

Why use Differential Privacy to Protect Data?

- The privacy loss as defined is guaranteed
  - Does not require unknowable assumptions about the attacker
- The amount of loss is quantifiable
  - In many cases it can be totaled across multiple releases and tracked like a "budget"

What are Drawbacks to Differential Privacy?



What are Drawbacks to Differential Privacy?

- Meeting the definition is hard for complex statistics/microdata
  - Counts/histograms are the most developed
- The meaning of the privacy loss is less intuitive
  - Compared with measures such as risk of re-identification
- No consensus on how to set the privacy parameter

Further Reading

- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis." In *Theory of cryptography conference*, pp. 265-284. Springer, Berlin, Heidelberg, 2006.
- Wasserman, Larry, and Shuheng Zhou. "A statistical framework for differential privacy." *Journal of the American Statistical Association* 105, no. 489 (2010): 375-389.
- Abowd, John M., and Lars Vilhuber. "How protective are synthetic data?." In *International Conference on Privacy in Statistical Databases*, pp. 239-246. Springer, Berlin, Heidelberg, 2008.
- Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. "Differential privacy: A primer for a non-technical audience." *Vand. J. Ent. & Tech. L.* 21 (2018): 209.
- Snoke, Joshua, and Claire McKay Bowen. "How statisticians should grapple with privacy in a changing data landscape." *Chance* 33, no. 4 (2020): 6-13.
- Snoke, Joshua, and Claire McKay Bowen. "Differential Privacy: What Is It?." *AMSTAT news: the membership magazine of the American Statistical Association* 501 (2019): 26-28.