# *Negotiating the privacy-accuracy trade-off in a differentially private release of earnings statistics*

**ISI WSC 2021: Formal privacy methods in NSO: Challenges and Solutions**

**Gerome Miklau**
Founder, Tumult Labs
Professor, Univ. of Massachusetts, Amherst

**Date: 7/15/2021**

About Us

# We are Tumult Labs

Founded in 2019 by leading experts in differential privacy

**Gerome Miklau**
**Founder**
Professor, Computer Science
UMass Amherst

**Michael Hay**
**Founder**
Professor, Computer Science
Colgate University

**Ashwin Machanavajjhala**
**Founder**
Professor, Computer Science
Duke University

**Helping organizations share and analyze sensitive data without compromising privacy**

**Leveraging groundbreaking differential privacy technology**

**At-scale deployments to enable data sharing**

United States® Census 2020

IRS

# Goals of this talk

- **Walk-through a real deployment of differential privacy:**
  - Releasing IRS income data to support **College Scorecard**
  - Emphasis on the lifecycle of a deployment: roles, responsibilities, interactions among parties.

TUMULT LABS

# Differential Privacy

A mathematically rigorous privacy standard that ensures *iron clad* privacy for individuals in the dataset.

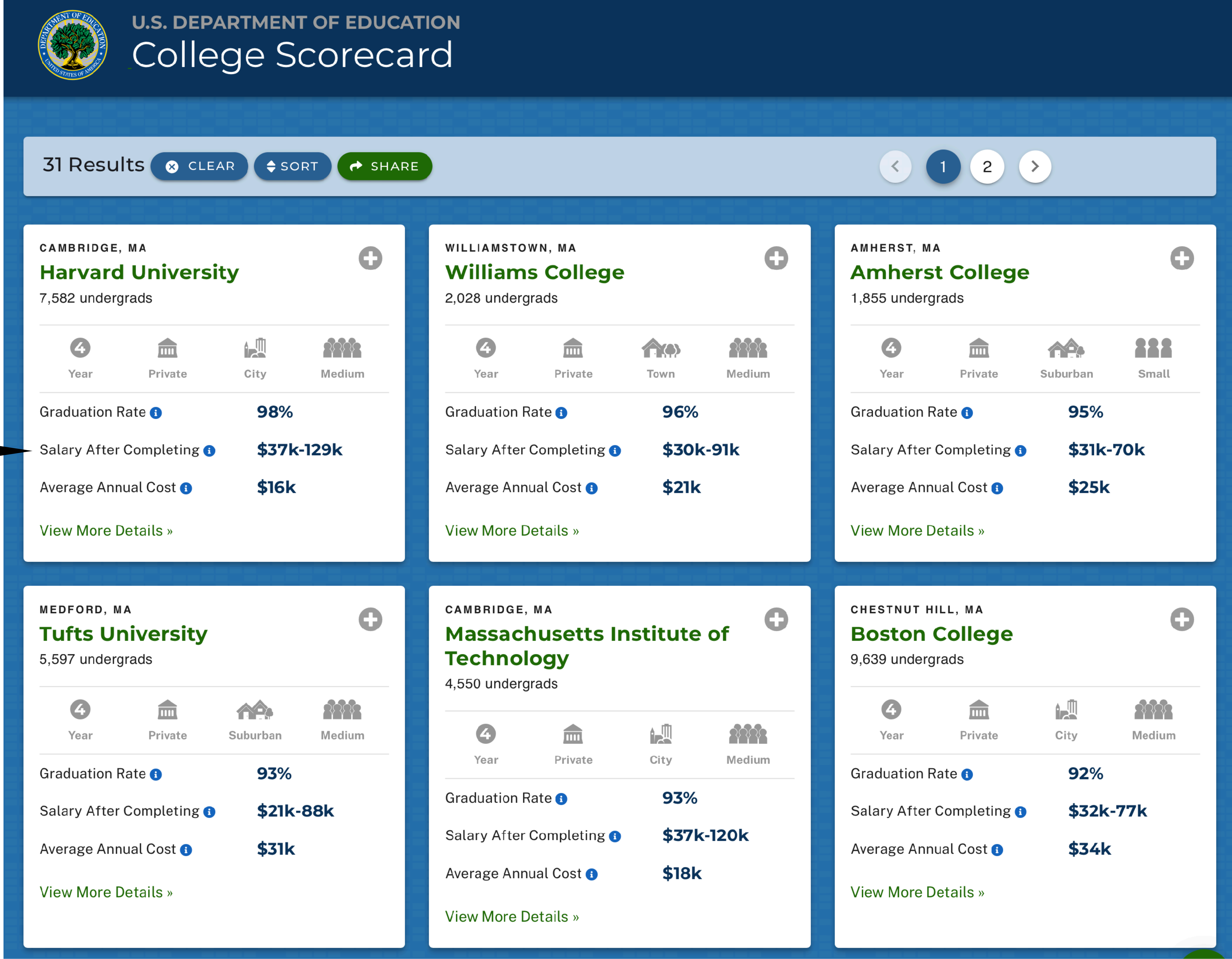
Outputs generated from differentially private algorithms are *future proof* from attacks that may not yet be invented.

A methodology that *meters privacy loss* even across multiple data releases.

## SEARCH RESULTS:

**State**: Massachusetts

median annual earnings of former students, one year after graduation



## The data owner

**IRS - SOI**

- The IRS is bound by law (U.S. Code Title 26) to protect all information provided on tax returns (even fact of filing)
- **Disclosure review board** must decide:
  - What to release
  - How to transform data to protect privacy

## The data analyst

**Dept. of Ed. (ED)**

- ED has access to educational records describing students and the degree programs they completed.
- ED asks the IRS to provide income data for the group of students in its sample, based on tax returns.

# Data owner challenges

- **Prior techniques were based primarily on suppression and ad hoc distortion of medians.**
  - The chosen suppression threshold led to > 70% of output rows unpublished!
- **Impossible to formally verify the privacy of this approach, especially when:**
  - the analyst requests increasingly detailed statistics,
  - an individual may appear more than once in the sensitive data,
  - releases are made annually,
  - SOI may release other summary statistics from the same source.

2020 and 2021: IRS has adopted differential privacy to address these concerns.

TUMULT LABS

# How differential privacy helped

- Rigorous, automated and quantifiable differential privacy guarantee helped simplify decision-making.
- Tumult's DP platform helped release more income statistics of students than previous releases (that used legacy SDL techniques) with comparable accuracy.
- Privacy algorithms can be made public without degrading the privacy guarantee.
- The released data currently powers the College Scorecard website.

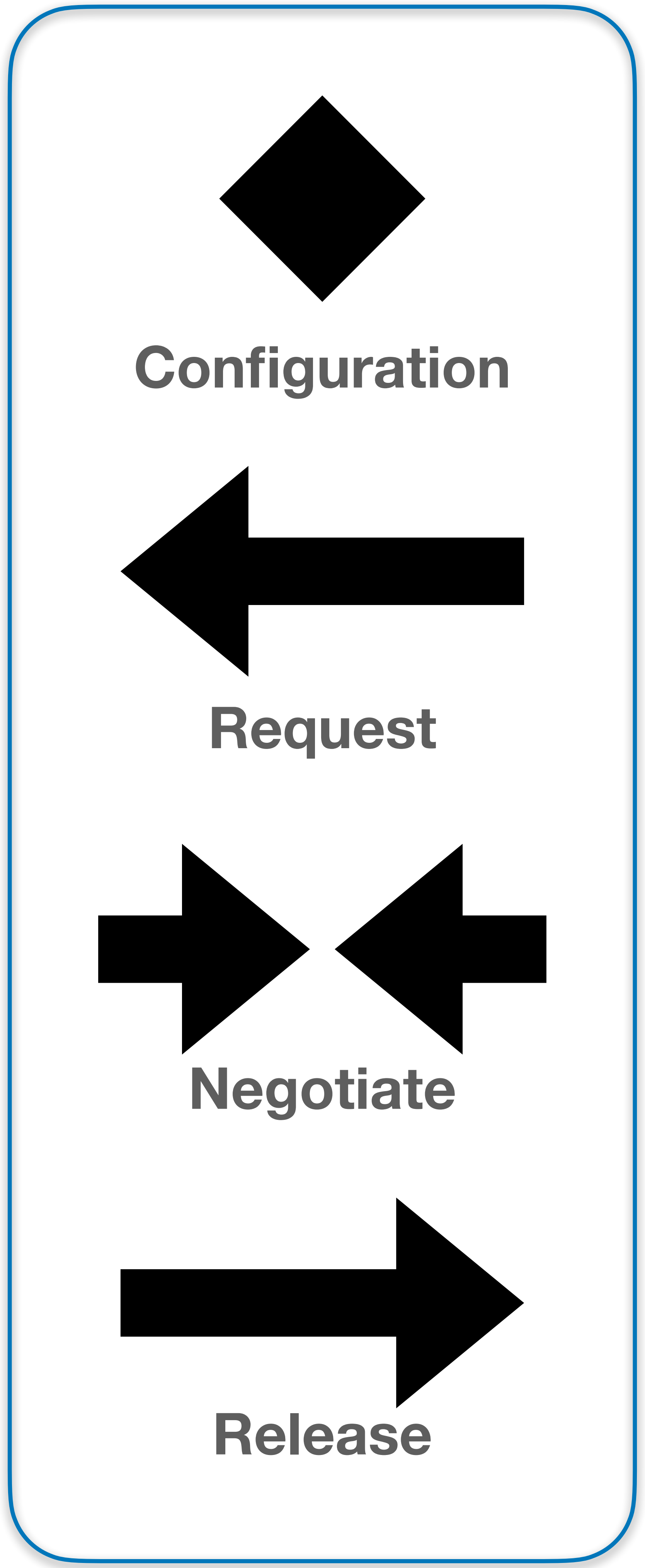
# The data owner

Dept. of Revenue

# The data analyst

Dept. of Schools

## DP deployment life-cycle



## The data owner

## The data analyst

Data owner defines data schema, domains, contributions of individuals.

**Configuration**

**Request**

Data analyst requests summary statistics desired to support their task.

Data owner selects epsilon & profiles achievable accuracy for analyst's request

**Negotiate**

Data analyst reviews accuracy; modifies request, clarifies priorities.

Data owner finalizes request and parameters; final execution to create release.

**Release**

# The analyst request: college scorecard

For each: COLLEGE, GENDER=any, PELL=any

### LINKED RECORDS REPORTING ON INDIVIDUALS

| SSN | COLLEGE | GENDER | PELL | INCOME |
|---|---|---|---|---|
| 131-43-12XX | 1002 | M | 0 | 94239 |
| 123-98-72XX | 1002 | F | 1 | 37481 |
| 148-68-24XX | 1003 | F | 0 | 54781 |
| 232-18-34XX | 1003 | M | 1 | 112963 |
| 514-98-32XX | 1002 | F | 0 | 29458 |
| 521-51-20XX | 1002 | F | 1 | 47158 |
| 542-14-86XX | 1003 | M | 0 | 39578 |
| 120-42-65XX | 1003 | F | 1 | 78415 |
| 194-85-36XX | 1007 | F | 0 | 68245 |
| 194-20-63XX | 1007 | F | 1 | 52489 |
| 352-58-84XX | 1008 | M | 0 | 48512 |

(This is not real data)

~ 10.2 million statistics in total

### SUMMARY STATISTICS

| COLLEGE | GENDER | PELL | COUNT | Q1 INCOME | Q2 INCOME | Q3 INCOME |
|---|---|---|---|---|---|---|
| 1001 | * | * | 201 | 34052 | 47634 | 56626 |
| 1002 | * | * | 145 | 40458 | 51479 | 59746 |
| 1003 | * | * | 9 | 34878 | 51282 | 61177 |
| 1004 | * | * | 11 | 42425 | 55762 | 66029 |
| … | … | … | … | … | … | … |
| 1001 | M | * | 65 | 41031 | 52084 | 63723 |
| 1002 | M | * | 321 | 37500 | 50687 | 58317 |
| 1003 | M | * | 297 | 36613 | 56250 | 58011 |
| 1004 | M | * | 134 | 35786 | 52809 | 65139 |
| … | … | … | … | … | … | … |
| 1007 | F | * | 12 | 39629 | 55180 | 66556 |
| 1007 | F | * | 9 | 35297 | 50140 | 63312 |

For each: COLLEGE, GENDER=M, PELL=any

TUMULT LABS

# DP algorithm design

SafeTables
TUMULT LABS

- SafeTables software:
  - **hardened implementation verified to satisfy the privacy guarantee**
  - **optimized accuracy (at any setting of epsilon)**
    - decompose request into optimally weighted measurements
    - combine measurements using inference
    - exploit constraints to improve error
  - **released statistics come with uncertainty measures**

**Major focus of DP research**

**Ektelo**
ACM SIGMOD 2018

**Matrix Mechanism & HDMM**
ACM PODS 2010
VLDB 2018

# Preparing for negotiation: accuracy profiling

- Data owner's disclosure review board (or privacy officers) determine maximum acceptable privacy loss bound (epsilon).
- For the requested workload and chosen epsilon, the data owner can profile the accuracy of the private output.

**COUNTS**

| Cell size | Error in Counts eps=1.0 | Error in Counts eps=5.0 |
|---|---|---|
| 1-10 | 2.773 | 0.555 |
| 11-20 | 2.561 | 0.512 |
| 21-40 | 2.527 | 0.505 |
| 41-80 | 2.633 | 0.527 |
| 81-160 | 2.593 | 0.519 |
| 161-300 | 2.615 | 0.523 |
| 300 | 2.614 | 0.523 |

**MEDIANS**

| Cell size | Error in Median(INCOME) eps=1.0 | Error in Median(INCOME) eps=5.0 |
|---|---|---|
| 1-10 | $34,136 | $22,709 |
| 11-20 | $28,434 | $10,563 |
| 21-40 | $23,494 | $7,256 |
| 41-80 | $15,227 | $3,568 |
| 81-160 | $8,398 | $1,724 |
| 161-300 | $4,505 | $909 |
| 300 | $1,569 | $341 |

# Many factors influence accuracy of DP:

- Overall epsilon privacy loss budget.
- The sophistication of the differentially private algorithm.
- Division of epsilon budget across sub-workloads (e.g. counts vs. quantiles)
- The requested output statistics
  - including statistics on overlapping populations.
- Properties of the input data:
  - distributional properties
  - size of cells

A major factor impacting accuracy of income quartiles

TUMULT LABS

# More breakouts, smaller cells

| OPEID | GENDER | PELL | Cell size | Number of cells |
|---|---|---|---|---|
| 6058 cells<br>defined by OPEID = x | | | 1-10 | 10 |
| | | | 11-20 | 531 |
| | | | 21-40 | 694 |
| | | | 41-80 | 908 |
| | | | 81-160 | 1003 |
| | | | 161-300 | 873 |
| | | | > 300 | 2039 |
| 12116 cells<br>defined by OPEID = x and **PELL = y** | | | 1-10 | 1117 |
| | | | 11-20 | 1431 |
| | | | 21-40 | 1777 |
| | | | 41-80 | 1950 |
| | | | 81-160 | 1942 |
| | | | 161-300 | 1467 |
| | | | > 300 | 2432 |
| 12116 cells<br>defined by OPEID = x and **GENDER = y** | | | 1-10 | 1135 |
| | | | 11-20 | 1401 |
| | | | 21-40 | 1781 |
| | | | 41-80 | 1957 |
| | | | 81-160 | 1951 |
| | | | 161-300 | 1468 |
| | | | > 300 | 2423 |

only ten small cells

a thousand small cells

TUMULT LABS

**Owner**

**Analyst**

At the epsilon we chose, you'll be seeing expected error of about 3 in counts, and $1500 to $15000 in quantile estimates.

That's fine for counts, but the **error is too great** for quantiles. Can you raise the epsilon?

No

But we could devote more of the privacy loss budget to quantile estimates: counts will have higher error of ~6, but quantiles will have lower error of $1000 to $10000.

SafeTables

That's better, but we care most about the error of medians (less about Q1 and Q3). And would it help if we dropped the GENDER breakdown?

. . .

Yes, if we focus on medians only, and drop the Gender breakdown, error will come down to between $500 and $5000.

Great! Let's do that.

# Final release

- After data owner and data analyst have agreed on workload, the disclosure review board completes final review.
- **Final release execution**:
    - Hardware-based secure random number generation is employed.
    - No retained seeds of random number generation.
- **Subsequent post-processing can be performed on output**
    - Suppress high-error data items to avoid mis-use.
    - This has no impact on privacy.

# Recap

Differential privacy provides a framework for sharing data while obeying regulatory constraints.

**The data owner gets** to impose a global bound on the privacy loss resulting from their release of data.

- This bound covers complex statistics and multiple releases.

**The data analyst gets** access to data they might not otherwise have.

- Any analyst request is possible, but a strict bound on privacy loss will limit accuracy or the range of statistics released.

**The role of software is**:

- to correctly guarantee epsilon-differential privacy
- to maximize accuracy of the release (for any epsilon)
- to support the privacy/accuracy negotiation

# Thank you.

# Questions?