# Building differentially private releases: basics of utility optimization

## Webinar 3 for Wikimedia Foundation, July 2022

**Damien Desfontaines**

# Recap from Webinar 2, outline of Webinar 3

- Differential privacy (DP) requires some changes to regular data analyses, and leads to a privacy-utility tradeoff
- Previously: using Tumult Analytics to run simple DP queries
- Today: how to optimize the trade-off, and get useful results

# Recap from Webinar 2, outline of Webinar 3

1. Solutions to homework exercises

2. Core insight: data size vs. relative error

3. Three hands-on exercises:
    - Splitting the privacy budget unevenly
    - Choosing good clamping bounds
    - Modifying the aggregation strategy

# Data size and relative error

What do we want about the output data?

Typically, we care about *relative* error: $$\frac{|real\_value - noisy\_value|}{real\_value}$$

# Data size and relative error

What do we want about the output data?

Typically, we care about *relative* error: $$\frac{|real\_value - noisy\_value|}{real\_value}$$

In simple cases, this is equivalent to: $$\frac{|noise|}{real\_value}$$

# Data size and relative error

What do we want about the output data?

Typically, we care about *relative* error: $$\frac{|real\_value - noisy\_value|}{real\_value}$$

In simple cases, this is equivalent to: $$\frac{|noise|}{real\_value}$$

Depends only on ε

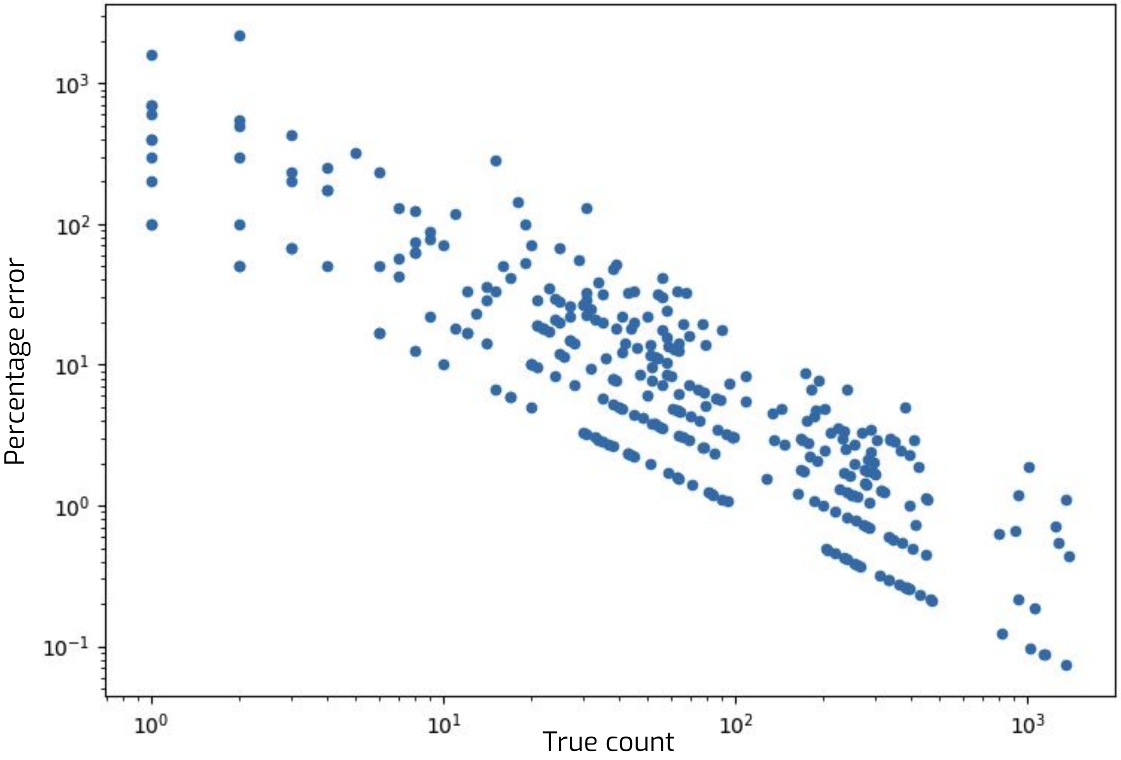
Depends on data size

# Data size and relative error

```
age_edu_keys = KeySet.from_dict(
  "age": list(range(5, 90)),
  "education_level": EDU_VALUES,
)

age_edu_query = (
  QueryBuilder("members")
  .groupby(age_edu_keys)
  .count()
)

result = session.evaluate(
  query,
  PureDPBudget(0.2),
)
```



# Exercise 1: Splitting the privacy budget

Three queries:

- Total count
- Count by age
- Count by age and gender

Goal: using a total budget of ε=3, getting the mean error of all three below 0.5%

# Exercise 1: Splitting the privacy budget

Hint 1: the noise magnitude for a counting query is on the order of $1/\varepsilon$.

Hint 2: we can look at the real values to get an idea of data magnitude.

- Total count (~54000)
- Count by age (median ~650, average ~600)
- Count by age and gender (median ~50, average ~160)

# Exercise 1: Splitting the privacy budget

Take-away: **more budget for finer aggregates**

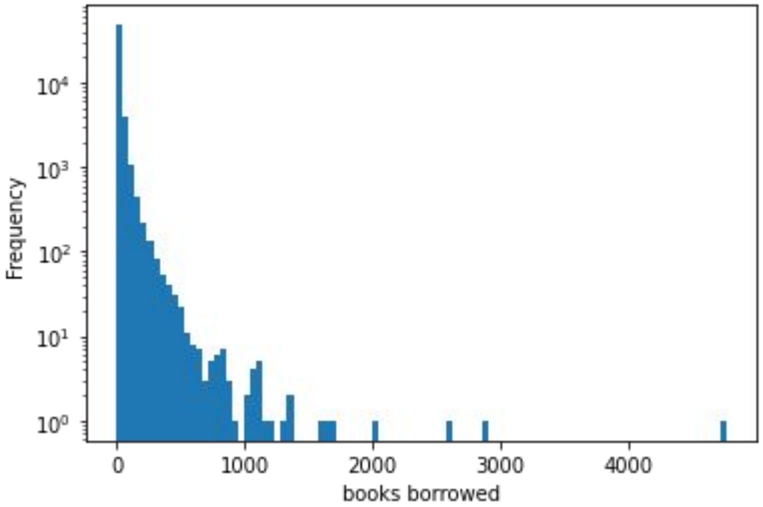
One possible solution:

- Total count: $\varepsilon = 0.01$
- Count by age: $\varepsilon = 0.49$
- Count by age and gender: $\varepsilon = 2.5$

# Exercise 2: Optimizing clamping bounds

Goal: publish the total number of books borrowed, by gender and age.

Main difficulty: clamping bounds?



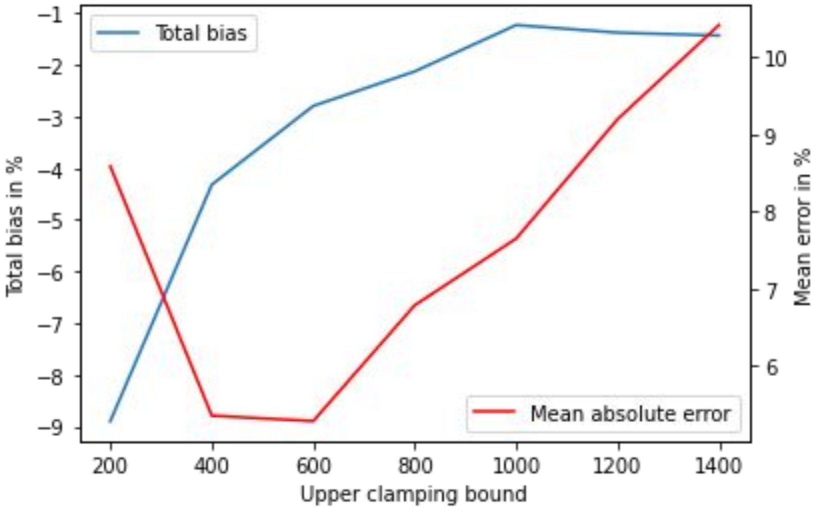
# Exercise 2: Optimizing clamping bounds

Takeaway: **error / bias trade-off**

A higher clamping bound means:

- less data loss: less bias
- more noise: more error

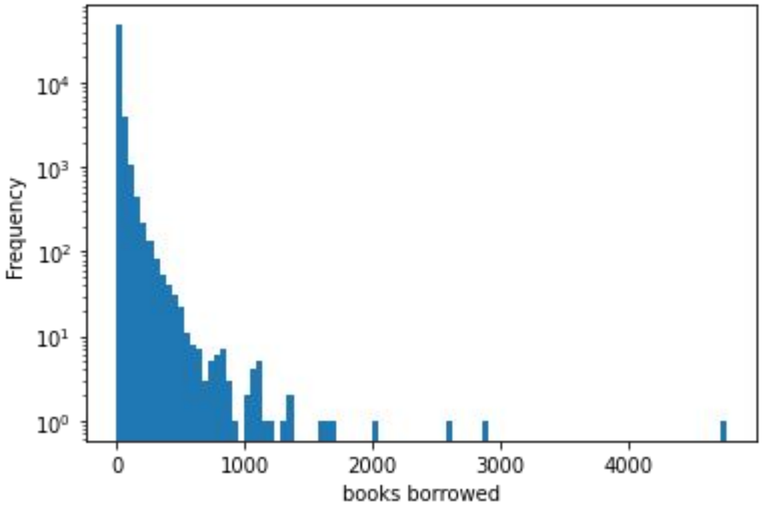
But when the clamping bound is too low, almost all the error comes from clamping.



# Exercise 3: Fine-tuning binning strategy

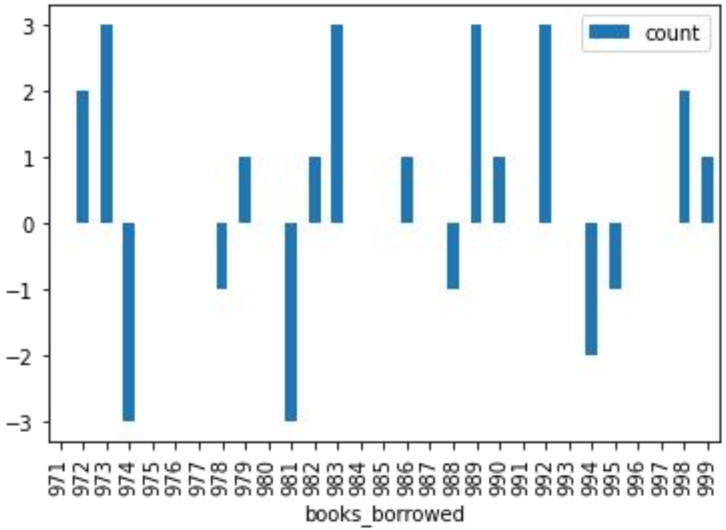
Goal: publish a histogram of number of books borrowed by library members

Main question: binning strategy?

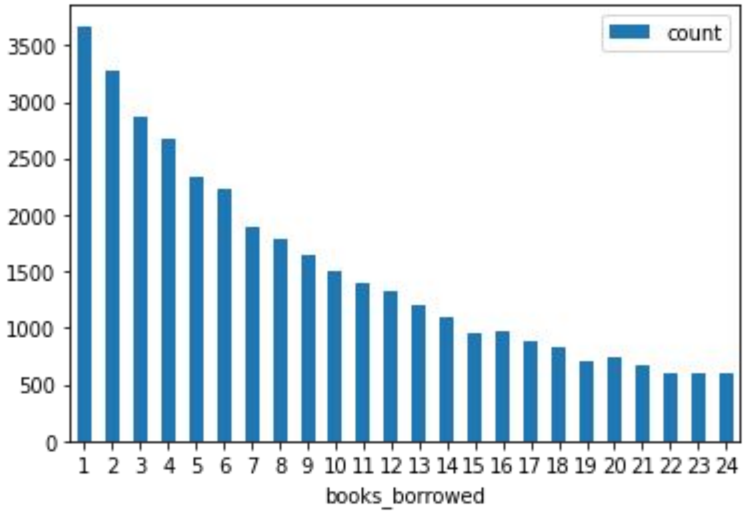


# Exercise 3: Fine-tuning binning strategy

Fine-grained:
pure noise for rare values



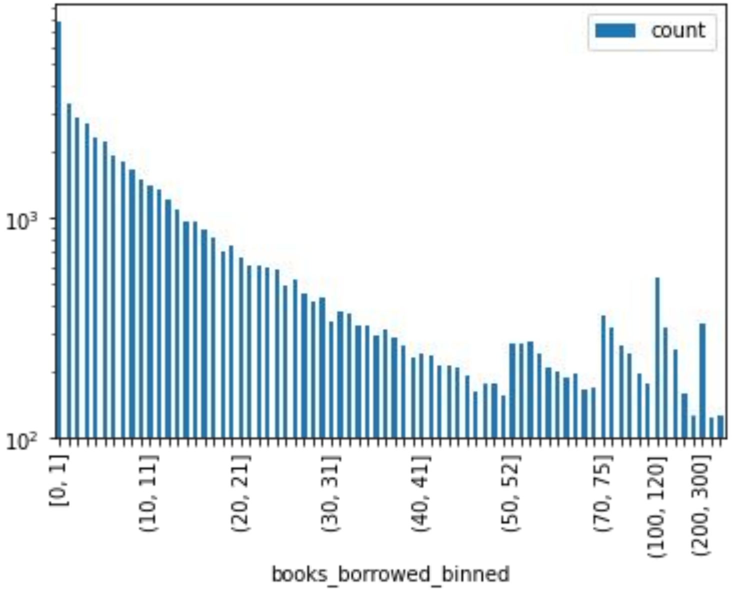

Coarse-grained:
loses data for frequent values



# Exercise 3: Fine-tuning the binning strategy

Takeaway: **larger bins for sparser data**

One possible “manual” strategy →




It’s also possible to do this in a DP way!

1. Use a small fraction of budget to determine binning strategy
2. Use the rest to compute counts

# Questions?

Damien Desfontaines
@TedOnPrivacy

tmlt.io/connect
tmlt.io/careers