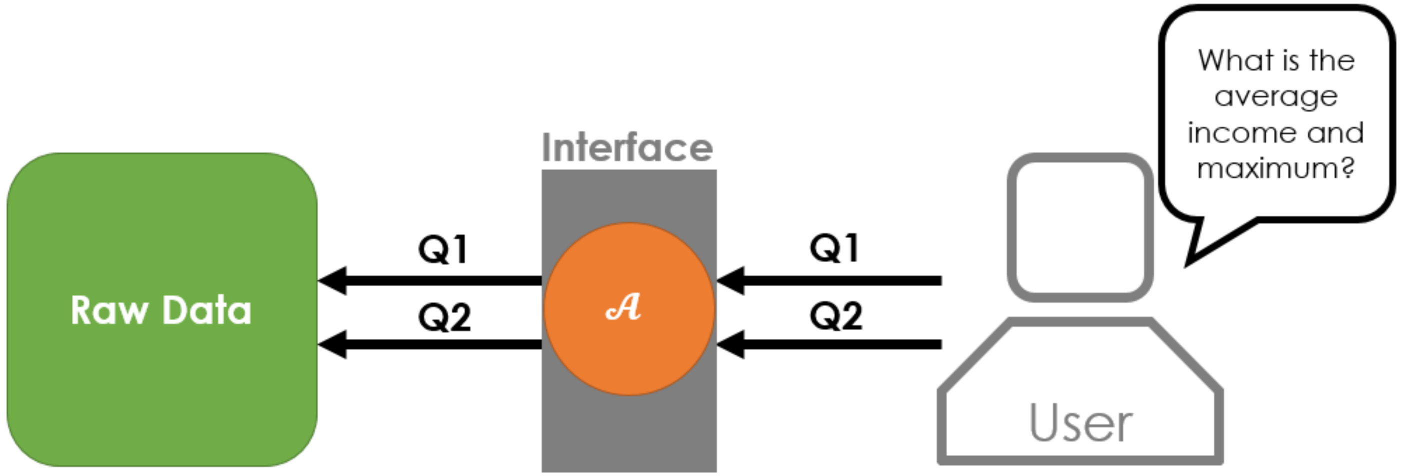
DATA PRIVACY AND ANONYMIZATION IN R

Sequential Composition

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Sequential Composition



- The privacy budget must be divided by two.

Male Fertility Data: Correction on Hours Sitting

```
# Mean and Variance of Hours Sitting
fertility %>%
    summarise_at(vars(Hours_Sitting), funs(mean, var))

# Apply the Laplace mechanism
set.seed(42)
rdoublex(1, 0.41, gs.mean / 0.1)
rdoublex(1, 0.19, gs.var / 0.1)
```

Male Fertility Data: Applying the Laplace mechanism

```
# Set Value of Epsilon
> eps <- 0.1 / 2

# GS of Mean and Variance
> gs.mean <- 0.01
> gs.var <- 0.01

# Apply the Laplace mechanism
> set.seed(42)
> rdoublex(1, 0.41, gs.mean / eps)
[1] 0.4496674

> rdoublex(1, 0.19, gs.var / eps)
[1] 0.2466982
```

For Hours Sitting in the Feritlity Data:

- GS Mean = 0.01
- GS Variance = 0.01
- Mean = 0.41
- Variance = 0.19

DATA PRIVACY AND ANONYMIZATION IN R

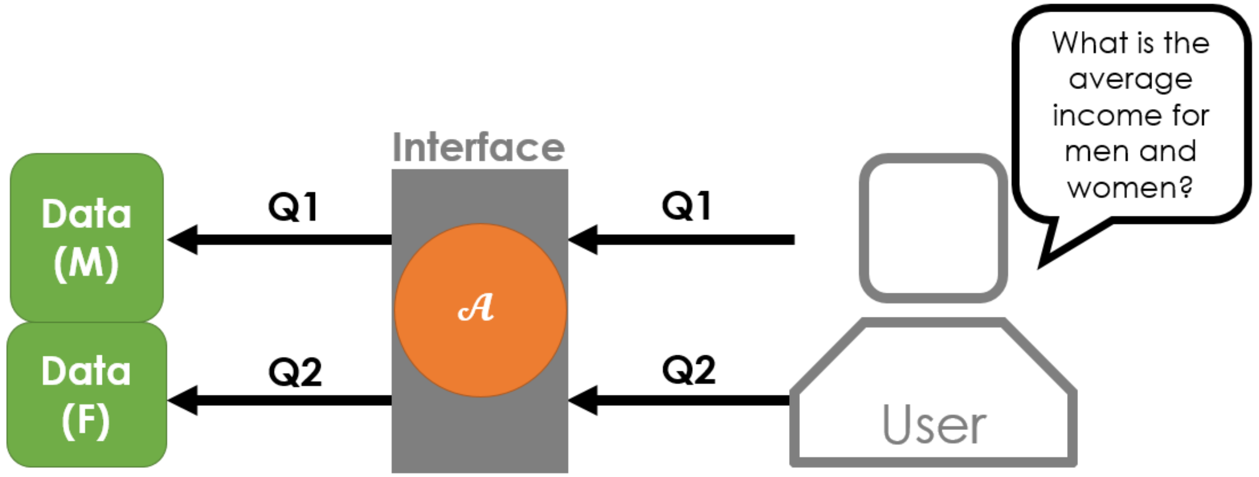
Let's practice!

DATA PRIVACY AND ANONYMIZATION IN R

Parallel Composition

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Parallel Composition



- The privacy budget does not need to be divided.
- The query with the most epsilon is the budget for the data.

Male Fertility Data: Prepping Data

```
# High_Fevers and Mean of Hours_Sitting
> fertility %>%
    filter(High_Fevers >= 0) %>%
    summarise_at(vars(Hours_Sitting), mean)
# A tibble: 1 x 1
  Hours_Sitting
          <dbl>
1     0.3932967

# No High_Fevers and Mean of Hours_Sitting
> fertility %>%
    filter(High_Fevers == -1) %>%
    summarise_at(vars(Hours_Sitting), mean)
# A tibble: 1 x 1
  Hours_Sitting
          <dbl>
1     0.5433333
```

Male Fertility Data: Applying Laplace mechanism

```
# Set Value of Epsilon
> eps <- 0.1

> # GS of mean for Hours_Sitting
> gs.mean <- 1 / 100

# Apply the Laplace mechanism
> set.seed(42)
> rdoublex(1, 0.39, gs.mean / eps)
[1] 0.4098337

> rdoublex(1, 0.54, gs.mean / eps)
[1] 0.5683491
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!

DATA PRIVACY AND ANONYMIZATION IN R

Post-processing

Claire McKay Bowen
Postdoctoral Researcher, Los Alamos National Laboratory

Male Fertility Data: Prepping Data

```
> fertility %>%
    count(Smoking)
# A tibble: 3 x 2
  Smoking Count
    <int> <int>
1      -1    56
2       0    23
3       1    21

# Set Value of Epsilon
> eps <- 0.1

# GS of Counts
> gs.count <- 1
```

Male Fertility Data: Applying the Laplace mechanism

```
# Apply the Laplace mechanism
> set.seed(42)
> smoking1 <- rdoublex(1, 56, gs.count / eps / 2) %>%
    round()
> smoking2 <- rdoublex(1, 23, gs.count / eps / 2) %>%
    round()

# Post-process based on previous queries
> smoking3 <- nrow(fertility) - smoking1 - smoking2

# Checking the noisy answers
> smoking1
[1] 60
> smoking2
[1] 29
> smoking3
[1] 11
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!

DATA PRIVACY AND ANONYMIZATION IN R

Impossible and Inconsistent Answers

Claire McKay Bowen

Postdoctoral Researcher, Los Alamos National Laboratory

Negative Counts: Prepping Data

```
# Set Value of Epsilon
> eps <- 0.01

# GS of counts
> gs.count <- 1

# Number of Participants with Abnormal Diagnosis
> fertility %>%
+     summarise_at(vars(Diagnosis), sum)
# A tibble: 1 x 1
  Diagnosis
      <int>
1        12
```

Negative Counts: Applying the Laplace mechanism

```
# Apply the Laplace mechanism and set.seed(22)
> set.seed(22)
> rdoublex(1, 12, gs.count / eps) %>%
    round()
[1] -79

# Apply the Laplace mechanism and set.seed(22)
> set.seed(22)
> rdoublex(1, 12, gs.count / eps) %>%
    round() %>%
    max(0)
[1] 0

# Suppose we set a different seed
> set.seed(12)
> noisy_answer <- rdoublex(1, 12, gs.count / eps) %>%
     round() %>%
     max(0)
> n <- nrow(fertility)
# ifelse example
> ifelse(noisy_answer > n, n, noisy_answer)
[1] 100
```

Normalizing Noise: Prepping Data

```
# Set Value of Epsilon
> eps <- 0.01

# GS of Counts
> gs.count <- 1

> fertility %>%
    count(Smoking)
# A tibble: 3 x 2
  Smoking Count
    <int> <int>
1      -1    56
2       0    23
3       1    21
```

Normalizing Noise: Applying the Laplace mechanism

```
# Apply the Laplace mechanism and set.seed(42)
> set.seed(42)
> smoking1 <- rdoublex(1, 56, gs.count / eps / 2) %>%
    max(0)
> smoking2 <- rdoublex(1, 23, gs.count / eps / 2) %>%
    max(0)
> smoking3 <- rdoublex(1, 21, gs.count / eps / 2) %>%
    max(0)

# Checking the noisy answers
> smoking <- c(smoking1, smoking2, smoking3)
> smoking
[1] 65.91684 37.17455  0.00000
```

Normalizing Noise: Constraining Results

```
# Normalize smoking
> normalized <- (smoking/sum(smoking)) * (nrow(fertility))
# Round the values
> round(normalized)
[1] 64 36  0
```

DATA PRIVACY AND ANONYMIZATION IN R

Let's practice!