A story about Local Differential Privacy

Hugo Richard

Senior Researcher - Criteo AI Lab

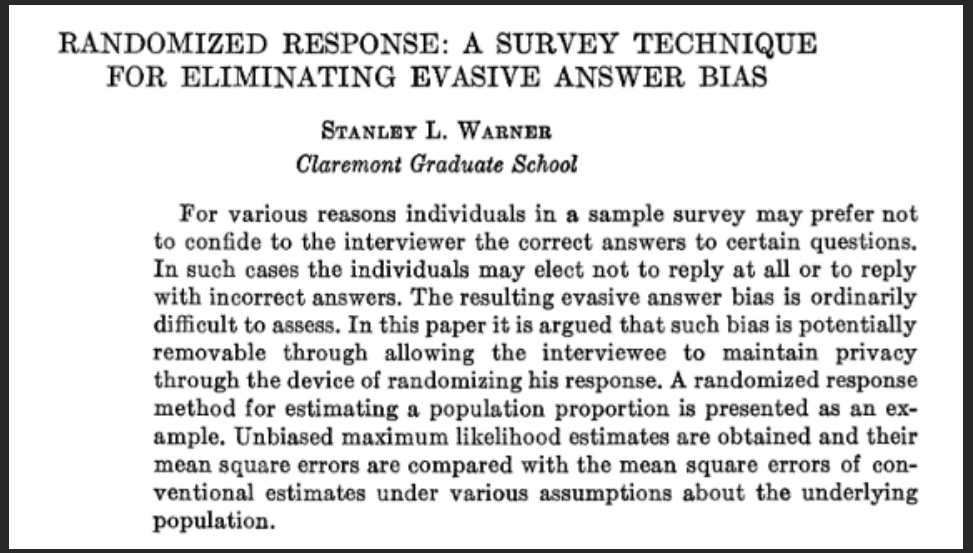
Part 1: Historical perspective and abortion in North Carolina

Part 2: New problems, new insights

Part 1: Historical perspective and abortion in North Carolina

Disclaimer: At Criteo, we do not collect such sensitive informations, the following is just an historical example of application of differential privacy.

Part 1: Historical perspective and abortion in North Carolina

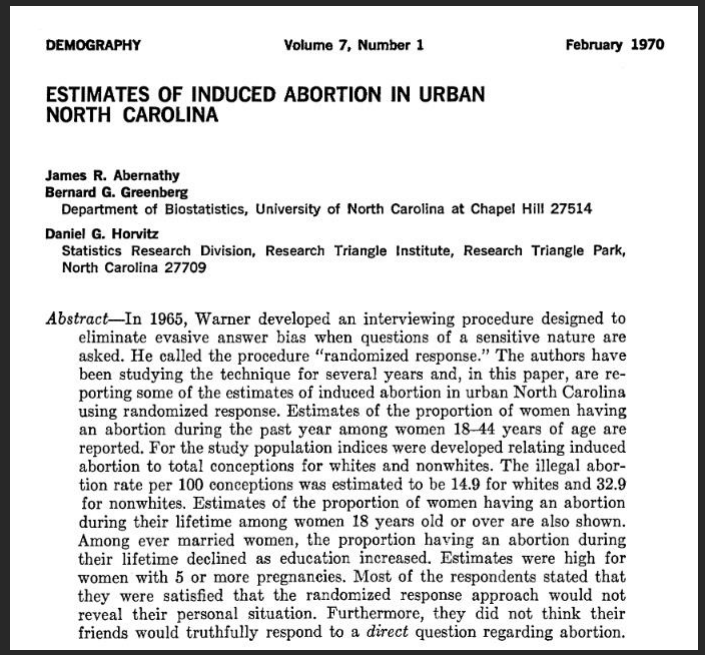


RANDOMIZED RESPONSE: A SURVEY TECHNIQUE FOR ELIMINATING EVASIVE ANSWER BIAS

STANLEY L. WARNER
Claremont Graduate School

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. In this paper it is argued that such bias is potentially removable through allowing the interviewee to maintain privacy through the device of randomizing his response. A randomized response method for estimating a population proportion is presented as an example. Unbiased maximum likelihood estimates are obtained and their mean square errors are compared with the mean square errors of conventional estimates under various assumptions about the underlying population.

1965



DEMOGRAPHY Volume 7, Number 1 February 1970

ESTIMATES OF INDUCED ABORTION IN URBAN NORTH CAROLINA

James R. Abernathy
Bernard G. Greenberg
Department of Biostatistics, University of North Carolina at Chapel Hill 27514
Daniel G. Horvitz
Statistics Research Division, Research Triangle Institute, Research Triangle Park, North Carolina 27709

Abstract—In 1965, Warner developed an interviewing procedure designed to eliminate evasive answer bias when questions of a sensitive nature are asked. He called the procedure "randomized response." The authors have been studying the technique for several years and, in this paper, are reporting some of the estimates of induced abortion in urban North Carolina using randomized response. Estimates of the proportion of women having an abortion during the past year among women 18–44 years of age are reported. For the study population indices were developed relating induced abortion to total conceptions for whites and nonwhites. The illegal abortion rate per 100 conceptions was estimated to be 14.9 for whites and 32.9 for nonwhites. Estimates of the proportion of women having an abortion during their lifetime among women 18 years old or over are also shown. Among ever married women, the proportion having an abortion during their lifetime declined as education increased. Estimates were high for women with 5 or more pregnancies. Most of the respondents stated that they were satisfied that the randomized response approach would not reveal their personal situation. Furthermore, they did not think their friends would truthfully respond to a *direct* question regarding abortion.

1970

Part 1: Historical perspective and abortion in North Carolina

RANDOMIZED RESPONSE: A SURVEY TECHNIQUE FOR ELIMINATING EVASIVE ANSWER BIAS

STANLEY L. WARNER
Claremont Graduate School

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. In this paper it is argued that such bias is potentially removable through allowing the interviewee to maintain privacy through the device of randomizing his response. A randomized response method for estimating a population proportion is presented as an example. Unbiased maximum likelihood estimates are obtained and their mean square errors are compared with the mean square errors of conventional estimates under various assumptions about the underlying population.

1965

DEMOGRAPHY Volume 7, Number 1 February 1970

ESTIMATES OF INDUCED ABORTION IN URBAN NORTH CAROLINA

James R. Abernathy
Bernard G. Greenberg
Department of Biostatistics, University of North Carolina at Chapel Hill 27514
Daniel G. Horvitz
Statistics Research Division, Research Triangle Institute, Research Triangle Park, North Carolina 27709

Abstract—In 1965, Warner developed an interviewing procedure designed to eliminate evasive answer bias when questions of a sensitive nature are asked. He called the procedure "randomized response." The authors have been studying the technique for several years and, in this paper, are reporting some of the estimates of induced abortion in urban North Carolina using randomized response. Estimates of the proportion of women having an abortion during the past year among women 18–44 years of age are reported. For the study population indices were developed relating induced abortion to total conceptions for whites and nonwhites. The illegal abortion rate per 100 conceptions was estimated to be 14.9 for whites and 32.9 for nonwhites. Estimates of the proportion of women having an abortion during their lifetime among women 18 years old or over are also shown. Among ever married women, the proportion having an abortion during their lifetime declined as education increased. Estimates were high for women with 5 or more pregnancies. Most of the respondents stated that they were satisfied that the randomized response approach would not reveal their personal situation. Furthermore, they did not think their friends would truthfully respond to a *direct* question regarding abortion.

1970

Estimates of induced abortion in urban North Carolina

Estimates of induced abortion in urban North Carolina

Race

○ White

○ Black

Estimates of induced abortion in urban North Carolina

Race	○ White	○ Black
Marital Status	○ Never married	○ Ever married

Estimates of induced abortion in urban North Carolina

- Race
  - White
  - Black
- Marital Status
  - Never married
  - Ever married
- Education (grade)
  - < 9th
  - 9th-12th
  - 13th and over

Estimates of induced abortion in urban North Carolina

Race	○ White		○ Black
Marital Status	○ Never married		○ Ever married
Education (grade)	○ < 9th	○ 9th-12th	○ 13th and over
Age	○ 18 - 31		○ 31 - 44

Estimates of induced abortion in urban North Carolina

Race	○ White		○ Black
Marital Status	○ Never married		○ Ever married
Education (grade)	○ < 9th	○ 9th-12th	○ 13th and over
Age	○ 18 - 31		○ 31 - 44
Number of pregnancies	○ 0-4		○ 5 and over

Estimates of induced abortion in urban North Carolina

Race	○ White		○ Black
Marital Status	○ Never married		○ Ever married
Education (grade)	○ < 9th	○ 9th-12th	○ 13th and over
Age	○ 18 - 31		○ 31 - 44
Number of pregnancies	○ 0-4		○ 5 and over
Abortion during past 12 months	○ Yes		○ No

Estimates of induced abortion in urban North Carolina

1970: Abortion is illegal and can lead to prosecutions

Abortion during past 12 months ○ Yes ○ No

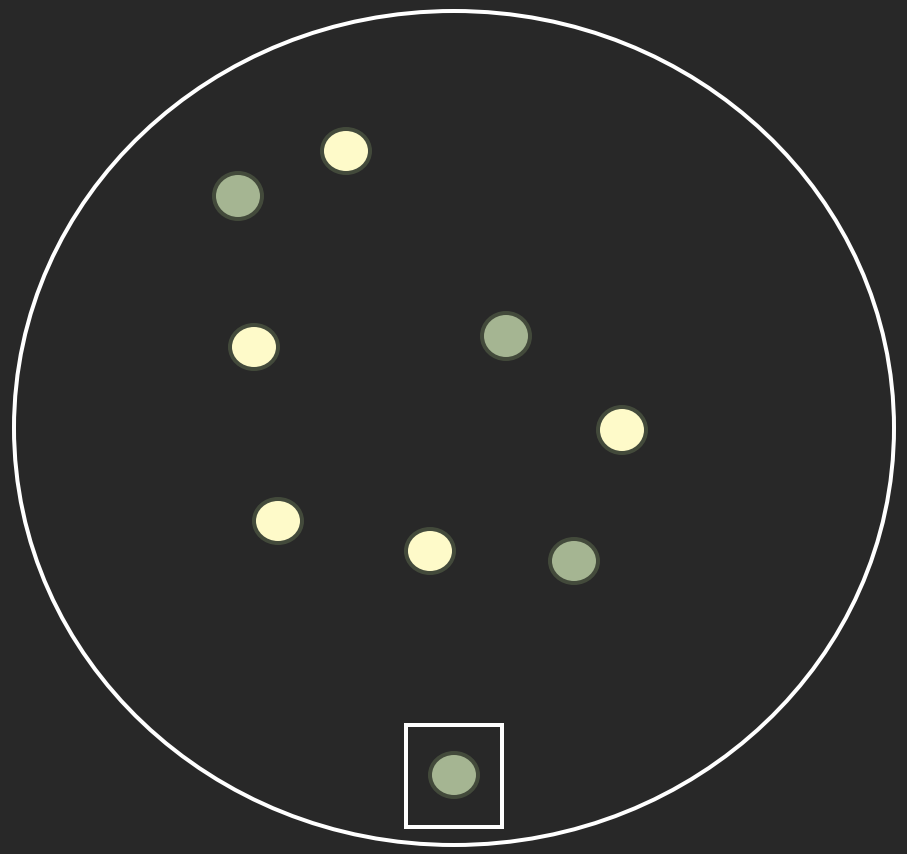
Estimates of induced abortion in urban North Carolina

Participation 3113 women were eligible (age, localization)

- 2.7 % Refused
- 92.7 % Accepted
- 5.1 % could not be located

Before knowing the experimental protocol

Estimates of induced abortion in urban North Carolina



- I was pregnant at some time during the past 12 months and had an abortion which ended the pregrancy
- I was born in the month of April

○ Yes ○ No

Estimates of induced abortion in urban North Carolina

Participants were asked …

… whether their friend would have answered truthfully to a direct question ?

17 % Yes

67 % No

16 % Undecided

Estimates of induced abortion in urban North Carolina

Participants were asked ...

... whether their friend would have answered truthfully to a direct question ?	17 % Yes	67 % No	16 % Undecided
... whether other people would think there was a trick to the box and that it is possible to figure out which question was answered ?	20 % Yes	60 % No	20 % Undecided

Estimates of induced abortion in urban North Carolina

Participants were asked ...

... whether their friend would have answered truthfully to a direct question ?	17 % Yes	67 % No	16 % Undecided
... whether other people would think there was a trick to the box and that it is possible to figure out which question was answered ?	20 % Yes	60 % No	20 % Undecided

What is your answer ?

Estimates of induced abortion in urban North Carolina

But unfortunately there was a trick …

Estimates of induced abortion in urban North Carolina

But unfortunately there was a trick …

What if I knew your birthday ?

known | - You are born in May
- You answered Yes → I know you had an abortion

Estimates of induced abortion in urban North Carolina

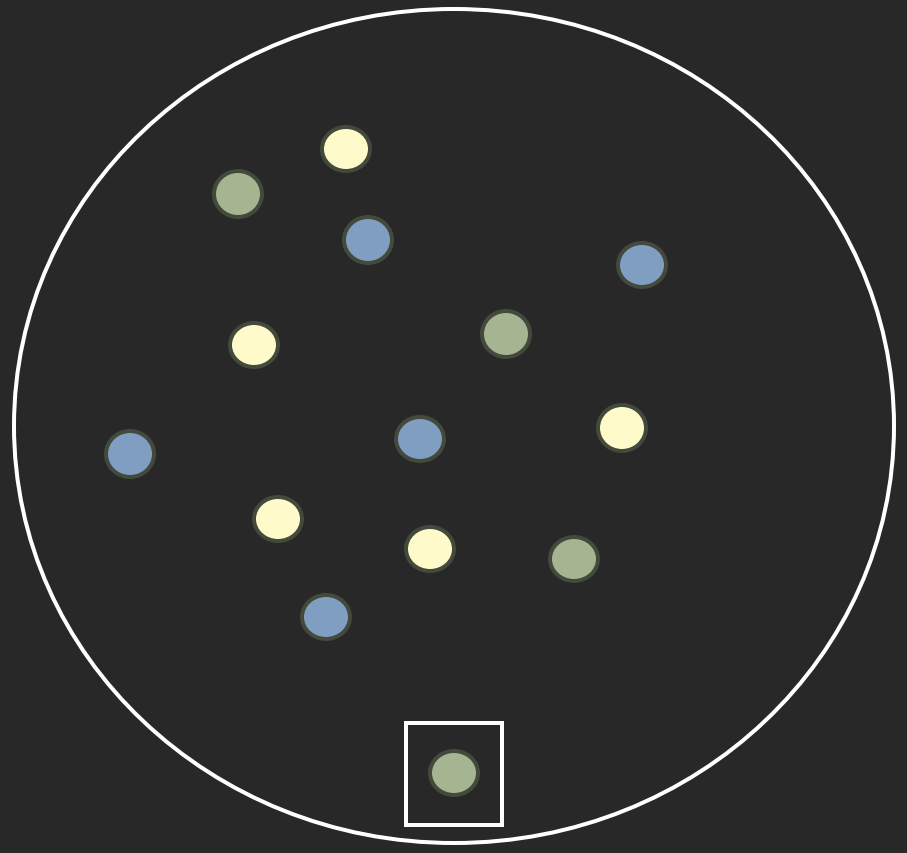
But unfortunately there was a trick …

What if I knew your birthday ?

known | - You are born in May
- You answered Yes → I know you had an abortion

Participating in the study is putting you at risk !

Randomized response: the correct way



- I was pregnant at some time during the past 12 months and had an abortion which ended the pregrancy (Abortion ball)
- Answer Yes (Yes ball)
- Answer No (No ball)

○ Yes ○ No

Measuring privacy with Local Differential Privacy

Measuring privacy with Local Differential Privacy

“Your answer only gives limited information about you”

Measuring privacy with Local Differential Privacy

"Your answer only gives limited information about you"

ε - Local Differential Privacy

For any sensitive informations s, s' ∈ {"abortion", "no abortion"} such that s ≠s'

For any possible answer a ∈ {"yes", "no"} it holds:

$$\frac{P(answer = a \mid sensitive\ information = s)}{P(answer = a \mid sensitive\ information = s')} \leq \exp(\varepsilon)$$

ε : privacy loss

Measuring privacy with Local Differential Privacy

"Your answer only gives limited information about you"

ε - Local Differential Privacy

For any sensitive informations s, s' ∈ {"abortion", "no abortion"} such that s ≠s'

For any possible answer a ∈ {"yes", "no"} it holds:

$$\frac{P(answer = a \mid sensitive\ information = s)}{P(answer = a \mid sensitive\ information = s')} \leq \exp(\varepsilon)$$

ε : privacy loss

Our mechanism guarantuees ε-local differential privacy if

$$\frac{P(Yes \mid Abortion)}{P(Yes \mid No\ abortion)}, \frac{P(Yes \mid No\ Abortion)}{P(Yes \mid Abortion)}, \frac{P(No \mid Abortion)}{P(No \mid No\ abortion)}, \frac{P(No \mid No\ Abortion)}{P(No \mid Abortion)}$$ are $\leq \exp(\varepsilon)$

Measuring privacy with Local Differential Privacy

$$\frac{P(Yes|\ Abortion)}{P(Yes\ |\ No\ abortion)}$$

Measuring privacy with Local Differential Privacy

$$\frac{P(Yes|\ Abortion)}{P(Yes\ |\ No\ abortion)}$$

If you had an Abortion, you answer Yes if you pick a Yes ball or you pick an Abortion ball
If you had No Abortion, you answer Yes if you pick a Yes ball

Measuring privacy with Local Differential Privacy

$$\frac{P(Yes|\ Abortion)}{P(Yes\ |\ No\ abortion)} = \frac{P(\ Picking\ a\ yes\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ yes\ ball)}$$

If you had an Abortion, you answer Yes if you pick a Yes ball or you pick an Abortion ball
If you had No Abortion, you answer Yes if you pick a Yes ball

Measuring privacy with Local Differential Privacy

$$\frac{P(Yes|\ Abortion)}{P(Yes\ |\ No\ abortion)} = \frac{P(\ Picking\ a\ yes\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ yes\ ball)}$$

$$\frac{P(No|\ No\ Abortion)}{P(No\ |\ Abortion)} = \frac{P(\ Picking\ a\ no\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ no\ ball)}$$

Measuring privacy with Local Differential Privacy

$$\frac{P(Yes|\ Abortion)}{P(Yes\ |\ No\ abortion)} = \frac{P(\ Picking\ a\ yes\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ yes\ ball)}$$

$$= \frac{20 + 0}{20} \leq \exp(0)$$

$$\frac{P(No|\ No\ Abortion)}{P(No\ |\ Abortion)} = \frac{P(\ Picking\ a\ no\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ no\ ball)}$$

$$= \frac{10 + 0}{10} \leq \exp(0)$$

Example 0 Abortion balls 20 Yes balls 10 No balls $\longrightarrow$ $\varepsilon = 0$ (maximum privacy)

But no one has answered the question about abortion !

Measuring privacy with Local Differential Privacy

$$\frac{P(Yes|\ Abortion)}{P(Yes\ |\ No\ abortion)} = \frac{P(\ Picking\ a\ yes\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ yes\ ball)}$$

$$= \frac{20 + 70}{20} \leq \exp(\ln(8))$$

$$\frac{P(No|\ No\ Abortion)}{P(No\ |\ Abortion)} = \frac{P(\ Picking\ a\ no\ ball) + P(Picking\ an\ abortion\ ball)}{P(Picking\ a\ no\ ball)}$$

$$= \frac{10 + 70}{10} \leq \exp(\ln(8))$$

Example 70 Abortion balls 20 Yes balls 10 No balls $\longrightarrow \varepsilon = \ln(8)$

"Most people" have answered the question but higher privacy loss

How do you treat these answers ?

"The lower the privacy loss (ε), the higher is the users' protection, the less precise your answers will be."

How do you treat these answers ?

"The lower the privacy loss (ε), the higher is the users' protection, the less precise your answers will be."

The statistician's perspective: you answer yes if you pick the yes ball or if you pick the abortion ball and had an abortion

How do you treat these answers ?

"The lower the privacy loss (ε), the higher is the users' protection, the less precise your answers will be."

The statistician's perspective: you answer yes if you pick the yes ball or if you pick the abortion ball and had an abortion

$$E\left[\frac{\#\ Yes\ answers}{\#\ answers}\right] = P(\text{pick the yes ball}) + P(\text{pick the abortion ball})\ \text{abortion rate}$$

How do you treat these answers ?

"The lower the privacy loss (ε), the higher is the users' protection, the less precise your answers will be."

The statistician's perspective: you answer yes if you pick the yes ball or if you pick the abortion ball and had an abortion

$$\mathrm{E}\left[\frac{\#\,Yes\ answers}{\#\ answers}\right] = \mathrm{P}(\text{pick the yes ball}) + \mathrm{P}(\text{pick the abortion ball})\ \text{abortion rate}$$

$$\widehat{\text{abortion}}\ \text{rate} = \frac{\frac{\#\,Yes\ answers}{\#\ answers} - P(pick\ the\ yes\ ball)}{P(pick\ the\ abortion\ ball)}$$

How do you treat these answers ?

"The lower the privacy loss (ε), the higher is the users' protection, the less precise your answers will be."

The statistician's perspective: you answer yes if you pick the yes ball or if you pick the abortion ball and had an abortion

$$\mathrm{E}\left[\frac{\#\,Yes\ answers}{\#\ answers}\right] = \mathrm{P}(\text{pick the yes ball}) + \mathrm{P}(\text{pick the abortion ball})\ \text{abortion rate}$$

$$\widehat{\text{abortion}}\ \text{rate} = \frac{\frac{\#\,Yes\ answers}{\#\ answers} - P(pick\ the\ yes\ ball)}{P(pick\ the\ abortion\ ball)}$$

Theorem (Warner, 1965).

$$\mathrm{E}[(\widehat{\text{abortion}}\ \text{rate} - \text{abortion rate})^2] \lesssim \min\left(\frac{1}{\varepsilon^2 n}, \frac{1}{n}\right)$$

Part 2: New problems, new insights

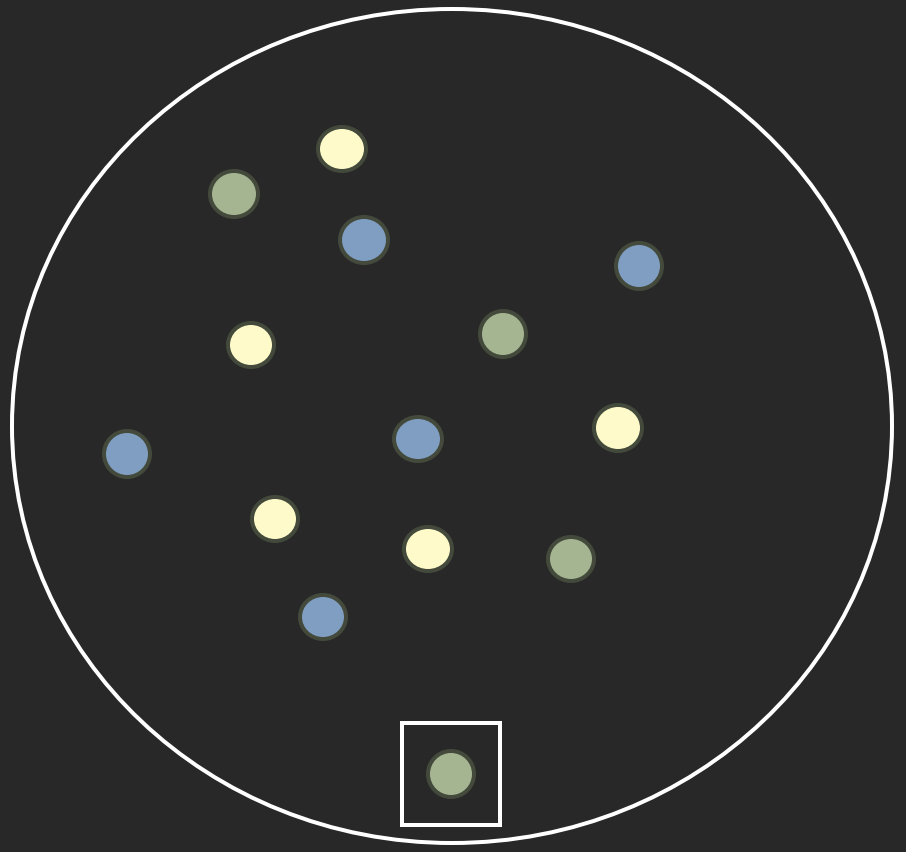
Part 2: New problems, new insights

Not all datasets are surveys

- Continuous, multi-dimensional data, multiple tasks
- Same user contributes multiple times

To illustrate the issues that may arise when a user contributes multiple time I will again use the abortion example (but again we do not collect that kind of data at Criteo)

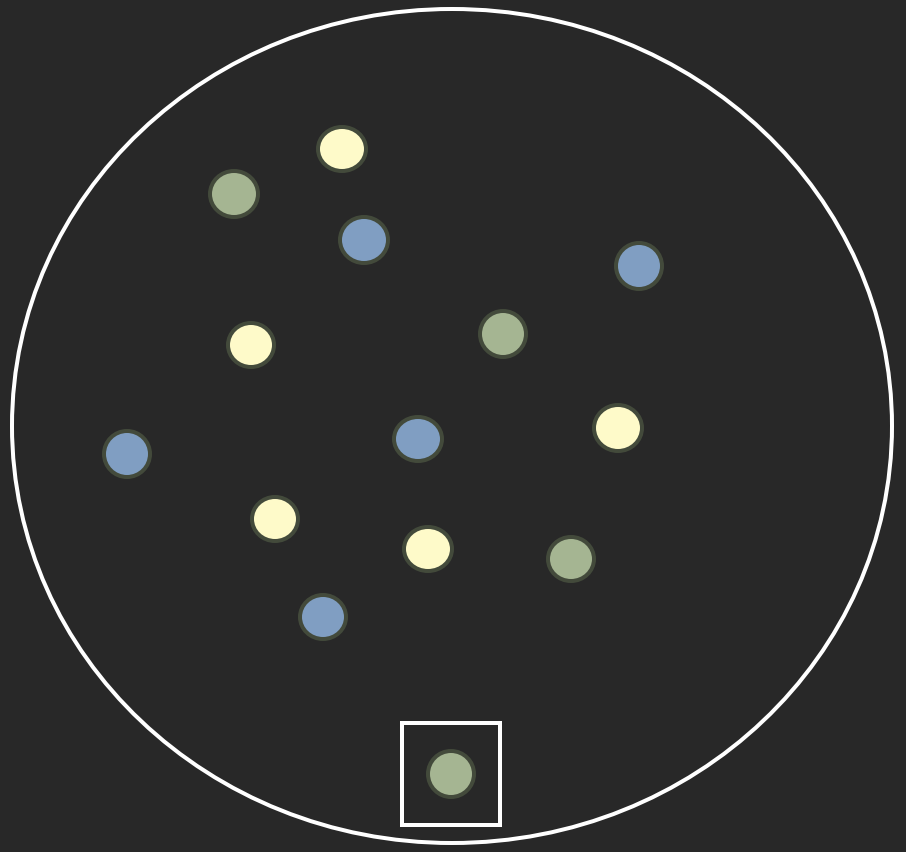
Part 2: New problems, new insights



- I was pregnant at some time during the past 12 months and had an abortion which ended the pregrancy (Abortion ball)
- Answer Yes (Yes ball)
- Answer No (No ball)

Assume that each user repeats that protocol m times

Part 2: New problems, new insights



- I was pregnant at some time during the past 12 months and had an abortion which ended the pregrancy (Abortion ball)
- Answer Yes (Yes ball)
- Answer No (No ball)

Assume that each user repeats that protocol m times
Give all answers to the statistician and repeat previous analysis.

Part 2: New problems, new insights

To get ε -LDP, I need to guarantuee that for any possible sequence of answers that the sensitive information does not matter too much.

Part 2: New problems, new insights

To get ε -LDP, I need to guarantuee that for any possible sequence of answers that the sensitive information does not matter too much.

In particular we should have:

$$\frac{P(Answer\ Yes\ m\ times|\ Abortion)}{P(Answer\ Yes\ m\ times|\ No\ abortion)} = \left(\frac{P(Answer\ Yes|\ Abortion)}{P(Answer\ Yes|\ No\ abortion)}\right)^{m} \leq \exp(\varepsilon)$$

Part 2: New problems, new insights

To get ε -LDP, I need to guarantuee that for any possible sequence of answers that the sensitive information does not matter too much.

In particular we should have:

$$\frac{P(Answer\ Yes\ m\ times|\ Abortion)}{P(Answer\ Yes\ m\ times|\ No\ abortion)} = \left(\frac{P(Answer\ Yes|\ Abortion)}{P(Answer\ Yes|\ No\ abortion)}\right)^{m} \leq \exp(\varepsilon)$$

Example 70 Abortion balls 20 Yes balls 10 No balls $\longrightarrow$ $\varepsilon = \ln(8)\, m$

Part 2: New problems, new insights

To get ε -LDP, I need to guarantuee that for any possible sequence of answers that the sensitive information does not matter too much.

In particular we should have:

$$\frac{P(Answer\ Yes\ m\ times|\ Abortion)}{P(Answer\ Yes\ m\ times|\ No\ abortion)} = \left(\frac{P(Answer\ Yes|\ Abortion)}{P(Answer\ Yes|\ No\ abortion)}\right)^{m} \leq \exp(\varepsilon)$$

Example 70 Abortion balls 20 Yes balls 10 No balls $\longrightarrow$ $\varepsilon = \ln(8)\, m$

There exists a better way than asking participants to reveal all their answers.
But this is a story for another time (see Corentin's poster at 6 PM today)

Conclusion

Take home message

- Local Differential privacy as middle ground between sharing and not sharing the data
- Vey strong notion of privacy as you do not trust the statistician
- Therefore, it is costly, you trade privacy against precision

Future work

Research-wise, many interesting questions around privacy and multiple interactions.

- Multidimensional data
- More complex models
- Correlated data

Thank you

Criteo AI Lab

Corentin Pla

Maxime Vono

Hugo Richard