CS2080: Applied Privacy for Data Science
Communicating Differential Privacy

School of Engineering & Applied Sciences
Harvard University

February 26, 2025

Housekeeping

1. HW 4 asks you to brainstorm ideas for the final project. Please brainstorm several ideas individually and hold off on forming teams until we’ve given you feedback.

2. We’ve posted an updated annotated course bibliography on the course webpage --- this may be a helpful resource as you brainstorm project ideas.

3. Interest in moving the HW deadlines to Fridays 11:59pm?

   We want to avoid cutting into your Friday evening plans.

Introduction

- Any given differential privacy (DP) deployment involves or impacts many parties:
  - Data curators (people or organizations collecting & managing data)
  - Data users (people analyzing or otherwise using privacy-protected data)
  - Data subjects (people contributing their data)
  - Developers & engineers (people implementing differentially-private mechanisms)
  - Policymakers (people deciding or proposing standards for how data should be protected)
  - etc.

Introduction

- Any given differential privacy (DP) deployment involves or impacts many parties:
  - Data curators (people or organizations collecting & managing data)
  - Data users (people analyzing or otherwise using privacy-protected data)
  - Data subjects (people contributing their data)
  - Developers & engineers (people implementing differentially-private mechanisms)
  - Policymakers (people deciding or proposing standards for how data should be protected)
  - etc.
- These parties are responsible for different sets of decisions, have different areas of expertise.
  - Data users have unique knowledge of how the data will be used downstream and can weigh in on accuracy needs
  - Data subjects must decide when to share data, and know how much privacy matters to them in specific contexts

Introduction

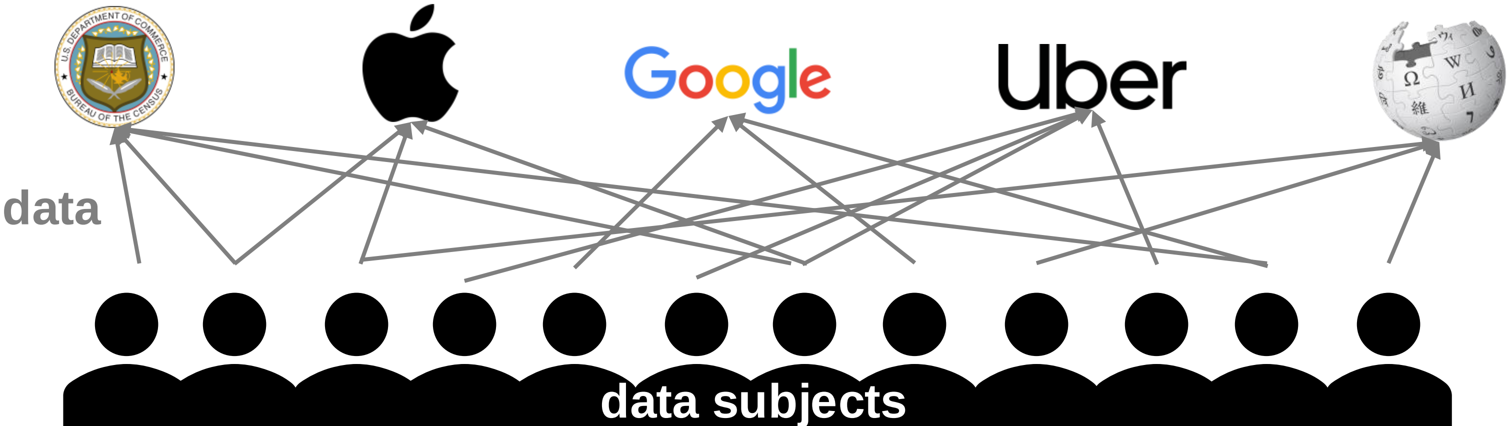
- Any given differential privacy (DP) deployment involves or impacts many parties:
  - Data curators (people or organizations collecting & managing data)
  - Data users (people analyzing or otherwise using privacy-protected data)
  - Data subjects (people contributing their data)
  - Developers & engineers (people implementing differentially-private mechanisms)
  - Policymakers (people deciding or proposing standards for how data should be protected)
  - etc.
- These parties are responsible for different sets of decisions, have different areas of expertise.
  - Data users have unique knowledge of how the data will be used downstream and can weigh in on accuracy needs
  - Data subjects must decide when to share data, and know how much privacy matters to them in specific contexts
- How can we ensure that each party is well-positioned to make informed decisions about DP?

Today’s goals

- Focus on communicating to data subjects
- By the end of today’s lecture you will be familiar with:
  - How DP has been explained to data subjects in practice, and how effective these approaches have been
  - Multiple *proposed* approaches for explaining DP that aim to improve upon current practices
  - Approaches to evaluating DP explanations with real people’s feedback

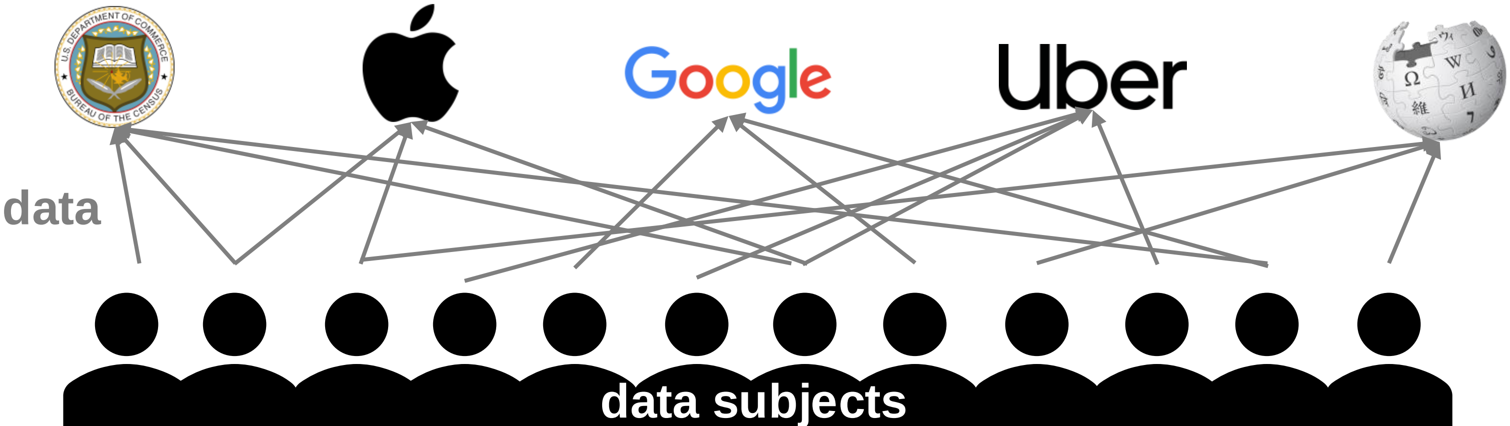
Who are data subjects?

- Subjects of the data
  - People whose data are used in analysis or to build models
- Examples in a DP context:
  - Respondents to the 2020 U.S. Census
  - Apple iPhone users
  - Google search users
  - Uber users
  - Wikipedia users (readers, editors)



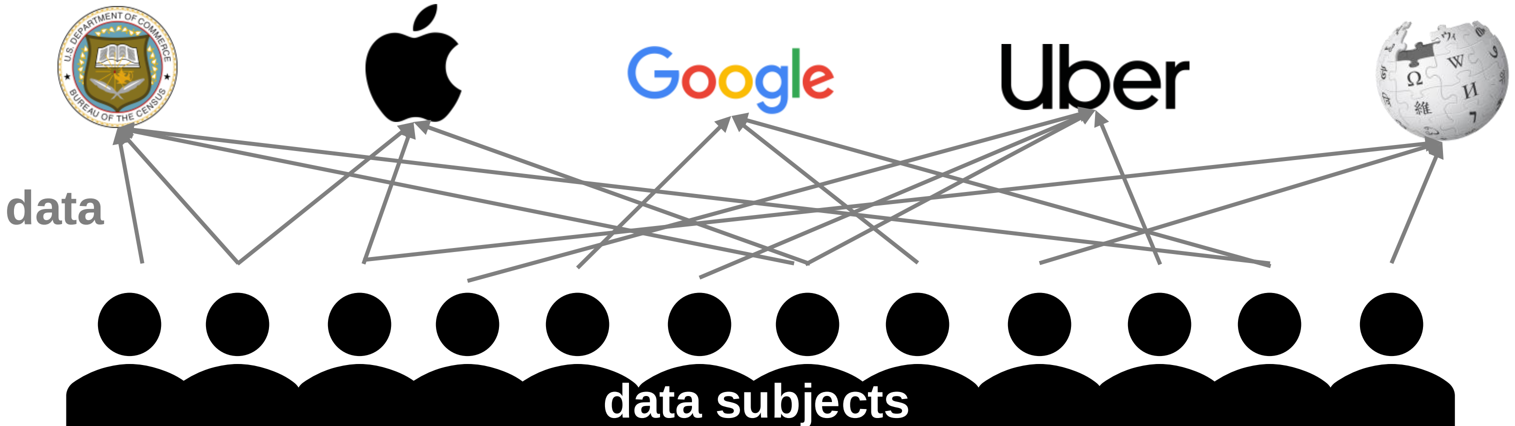
Who are data subjects?

- In most contexts, cannot assume computer science, math, or statistics background
- May not have time or interest in investing lots of time into understanding DP
- May have differing views on what privacy means or “how much” privacy they want
- Must decide whether to share data vs. not share data, or provide truthful vs. untruthful data



How might this information inform the type of explanations we develop?

- Present information without jargon and when presenting numeric information, do not assume mathematical background
- Explanations should not take more than a minute or two to understand
- Convey the strength of privacy protections and should align DP's guarantees with common understandings of privacy
- Help data subjects with deciding whether to share data vs. not share data

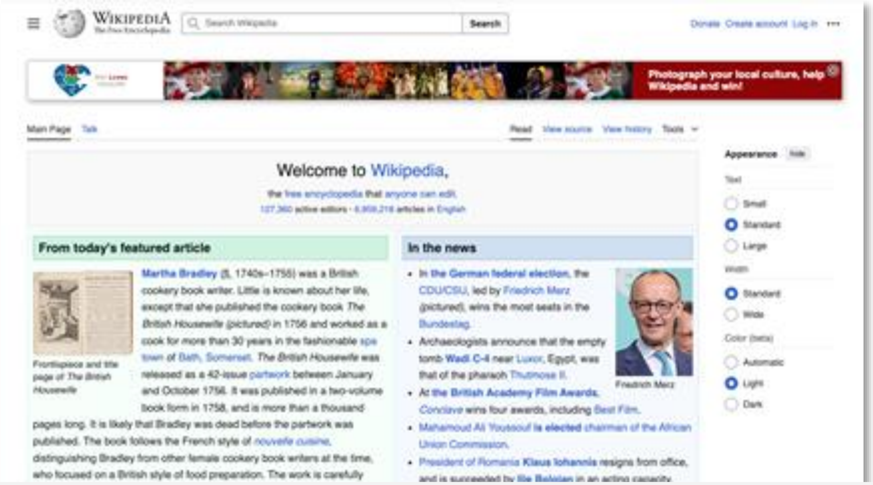


Discussion

Recall the Wikimedia Foundation's deployment of differential privacy to publish counts of pageviews by country. How would you explain DP's guarantees to a typical Wikipedia reader? Imagine this explanation will be posted on Wikipedia's homepage.

Some things to think about:

- What technical information is important to include vs. omit?
- How might you incorporate visuals into your explanation?
- How much time would a reader need to digest your explanation?
- How would you rigorously test whether readers understood the explanation?



Welcome to Wikipedia,

the free encyclopedia that anyone can edit.

127,360 active editors · 6,958,218 articles in English

From today's featured article

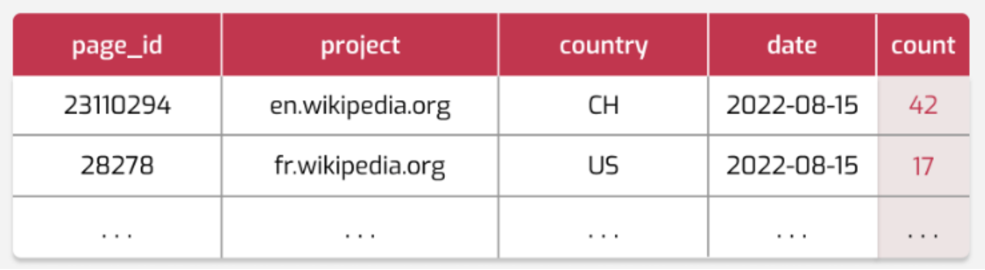
Frontispiece and title page of *The British Housewife*

Martha Bradley (fl. 1740s–1755) was a British cookery book writer. Little is known about her life, except that she published the cookery book *The British Housewife* (pictured) in 1756 and worked as a cook for more than 30 years in the fashionable spa town of Bath, Somerset. *The British Housewife* was released as a 42-issue partwork between January and October 1756. It was published in a two-volume book form in 1758, and is more than a thousand pages long. It is likely that Bradley was dead before the partwork was published. The book follows the French style of nouvelle cuisine, distinguishing Bradley from other female cookery book writers at the time, who focused on a British style of food preparation. The work is carefully

In the news

- In **the German federal election**, the CDU/CSU, led by Friedrich Merz (pictured), wins the most seats in the Bundestag.
- Archaeologists announce that the empty tomb **Wadi C-4** near Luxor, Egypt, was that of the pharaoh Thutmose II.
- At **the British Academy Film Awards**, Conclave wins four awards, including Best Film.
- Mahamoud Ali Youssouf **is elected** chairman of the African Union Commission.
- President of Romania **Klaus Iohannis** resigns from office, and is succeeded by **Ilie Bolojan** in an acting capacity.

Friedrich Merz



page_id	project	country	date	count
23110294	en.wikipedia.org	CH	2022-08-15	42
28278	fr.wikipedia.org	US	2022-08-15	17
...	...	...	...	...

https://www.tmlt.io/resources/publishing-wikipedia-usage-data-with-strong-privacy-guarantees

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
  - Unsubstantial
  - Techniques
  - Enables
  - Trust
  - Risk
  - Technical

Cummings, Kaptchuk, Redmiles 2021

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
  - Unsubstantial
    - “Differential privacy is the gold standard in data privacy and protection and is widely recognized as the strongest guarantee of privacy available.”
  - Techniques
  - Enables
  - Trust
  - Risk
  - Technical

Cummings, Kaptchuk, Redmiles 2021

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
  - Unsubstantial
  - Techniques
    - “Differential Privacy injects statistical noise into collected data in a way that protects privacy without significantly changing conclusions.”
  - Enables
  - Trust
  - Risk
  - Technical

Cummings, Kaptchuk, Redmiles 2021

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
    - Unsubstantial
    - Techniques
    - Enables
        - “Differential Privacy allows analysts to learn useful information from large amounts of data without compromising an individual’s privacy.”
    - Trust
    - Risk
    - Technical

Cummings, Kaptchuk, Redmiles 2021

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
  - Unsubstantial
  - Techniques
  - Enables
  - Trust
    - “Differential privacy is a novel, mathematical technique to preserve privacy which is used by companies like Apple and Uber.”
  - Risk
  - Technical

Cummings, Kaptchuk, Redmiles 2021

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
  - Unsubstantial
  - Techniques
  - Enables
  - Trust
  - Risk
    - “Differential privacy protects a user’s identity and the specifics of their data, meaning individuals incur almost no risk by joining the dataset.”
  - Technical

Cummings, Kaptchuk, Redmiles 2021

What’s happening in practice?

- In late 2019, researchers analyzed **76 explanations of DP** from industry (e.g., Google, Apple, Microsoft, Uber, start-ups, an investment firm), media outlets, and the academic literature
- They found six themes:
  - Unsubstantial
  - Techniques
  - Enables
  - Trust
  - Risk
  - Technical
    - “Differential privacy ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis. It follows that no risk is incurred by joining the database, providing a mathematically rigorous means of coping with the fact that distributional information may be disclosive.” (Dwork 08)

Cummings, Kaptchuk, Redmiles 2021

How well do these explanations set privacy expectations?

- Ideally, explanations of DP set accurate privacy expectations. That is, data subjects know what could and could not happen as a result of sharing their data.

- How might we test how well the previous explanations set accurate privacy expectations?
  - Run a study with real people, show them the explanations, and ask them questions about what might happen as a result of data sharing

How well do these explanations set privacy expectations?

Imagine that you work in the banking industry. You are friends with a group of other people who work in banking companies in your city. One of your friends is part of a transparency initiative that is trying to publish general statistics about pay in the banking industry. As part of this initiative, they have asked everyone in the group to share their salaries and job titles using an online web form on the initiative's website.

An explanation of DP or no mention of privacy protections

Banking and medical scenario tested across 1,208 participants recruited from Amazon Mechanical Turk (MTurk). Banking scenario shown above.

Cummings, Kaptchuk, Redmiles 2021

How well do these explanations set privacy expectations?

T/F questions about privacy expectations:

- **Criminal or foreign gov** that hacks the initiative could learn my salary/job title
- **Law enforcement** with a court order could access my salary/job title
- **Friend** collecting data will not learn be able to my salary/job title
- **Data analyst** working on the initiative could learn my exact salary/job title
- **Graphs or charts** made info given to the initiative could reveal my salary/job title
- **Data** that the initiative shares with other organizations doing salary research could reveal my salary/job title

Cummings, Kaptchuk, Redmiles 2021

How well do these explanations set privacy expectations?

T/F questions about privacy expectations:

- **Criminal or foreign gov** that hacks the initiative could learn my salary/job title
- **Law enforcement** with a court order could access my salary/job title
- **Friend** collecting data will not be able to learn to my salary/job title
- **Data analyst** working on the initiative could learn my exact salary/job title
- **Graphs or charts** made info given to the initiative could reveal my salary/job title
- **Data** that the initiative shares with other organizations doing salary research could reveal my salary/job title

Assuming the central model, with a "typical" deployment and small values of epsilon

Cummings, Kaptchuk, Redmiles 2021

How well do these explanations set privacy expectations?

T/F questions about privacy expectations:

- **Criminal or foreign gov** that hacks the initiative could learn my salary/job title
- **Law enforcement** with a court order could access my salary/job title
- **Friend** collecting data will not be able to learn to my salary/job title
- **Data analyst** working on the initiative could learn my exact salary/job title
- **Graphs or charts** made info given to the initiative could reveal my salary/job title
- **Data** that the initiative shares with other organizations doing salary research could reveal my salary/job title

For each explanation, participants answered correctly roughly half the time --- **about as good as random guessing**.

Assuming the central model, with a "typical" deployment and small values of epsilon

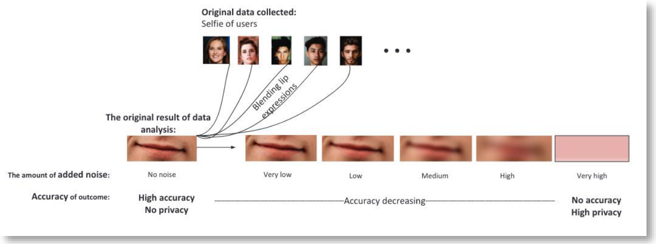
Cummings, Kaptchuk, Redmiles 2021

There are several *proposed* approaches to explaining DP to data subjects…

> "To respect your personal information privacy and ensure best user experience, the data shared with the app will be processed via the differential privacy (DP) technique. That is, the app company will store your data but only use the aggregated statistics with modification so that your personal information cannot be learned. However, your personal information may be leaked if the company's database is compromised." [Xiong, Wang, Li, Jha '20]

TEXT DESCRIPTIONS

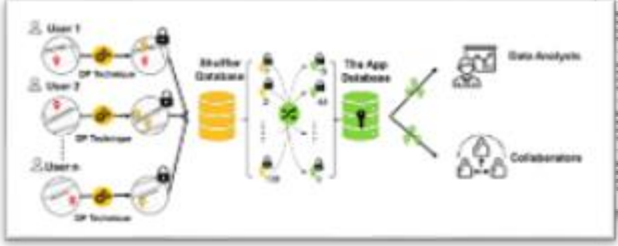
(e.g., Xiong et al. 2020, Cummings et al. 2021, Smart et al. 2023, Franzen et al. 2023)



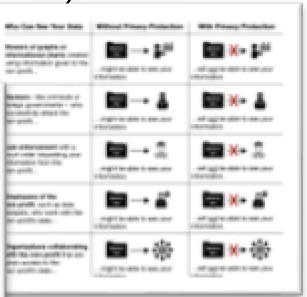
[Karegar Alaqra Fischer-Hübner 2022]

METAPHORS

(e.g., Bullek et al. 2017, Karegar et al. 2022)



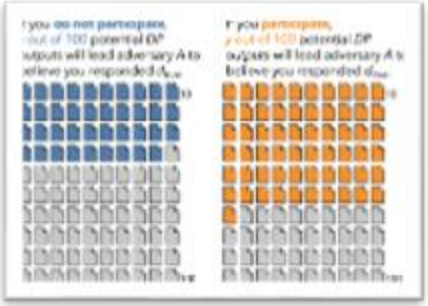
(Xiong, Wu, Wang, Proctor, Blocki, Li, Jha '22)


Who Can See Your Data
Without Privacy Protection
With Privacy Protection

[Smart, Nanayakkara, Cummings, Kaptchuk, Redmiles '24]

DIAGRAMS & TABLES

(e.g., Bullek et al. 2017, Karegar et al. 2022, Smart et al. 2024, Xiong et al. 2022, Wen et al. 2023)

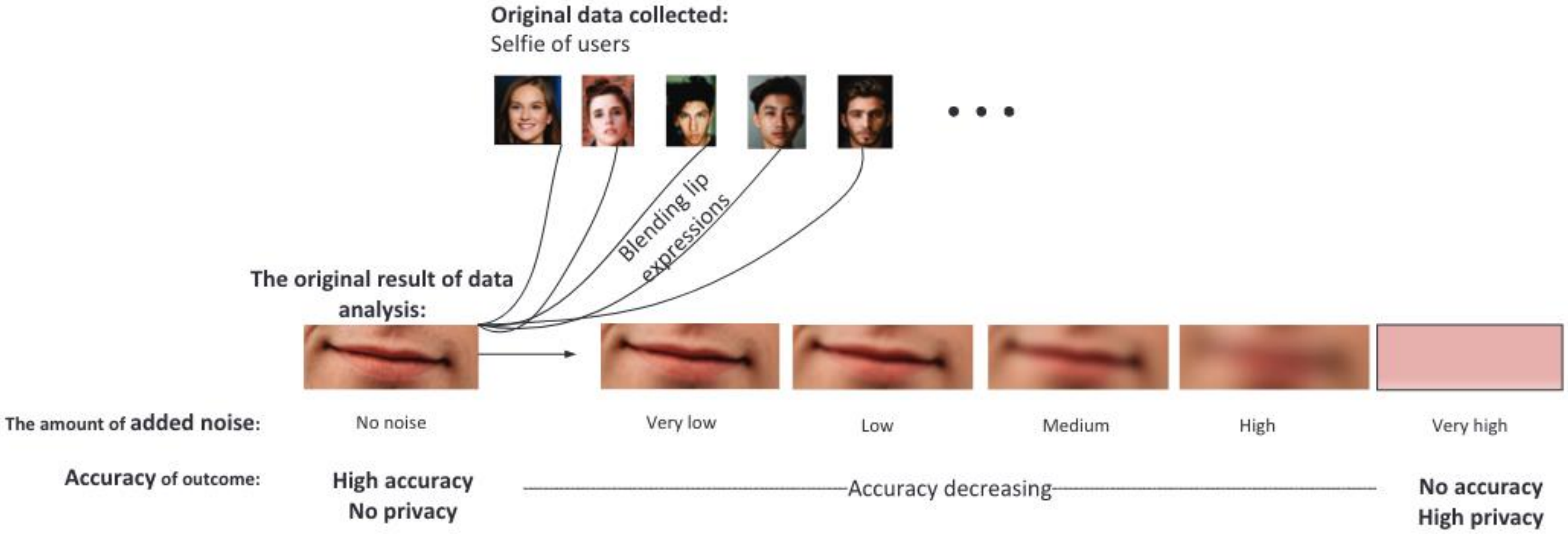


[Nanayakkara, Smart, Cummings, Kaptchuk, Redmiles '23]

VISUALIZATIONS

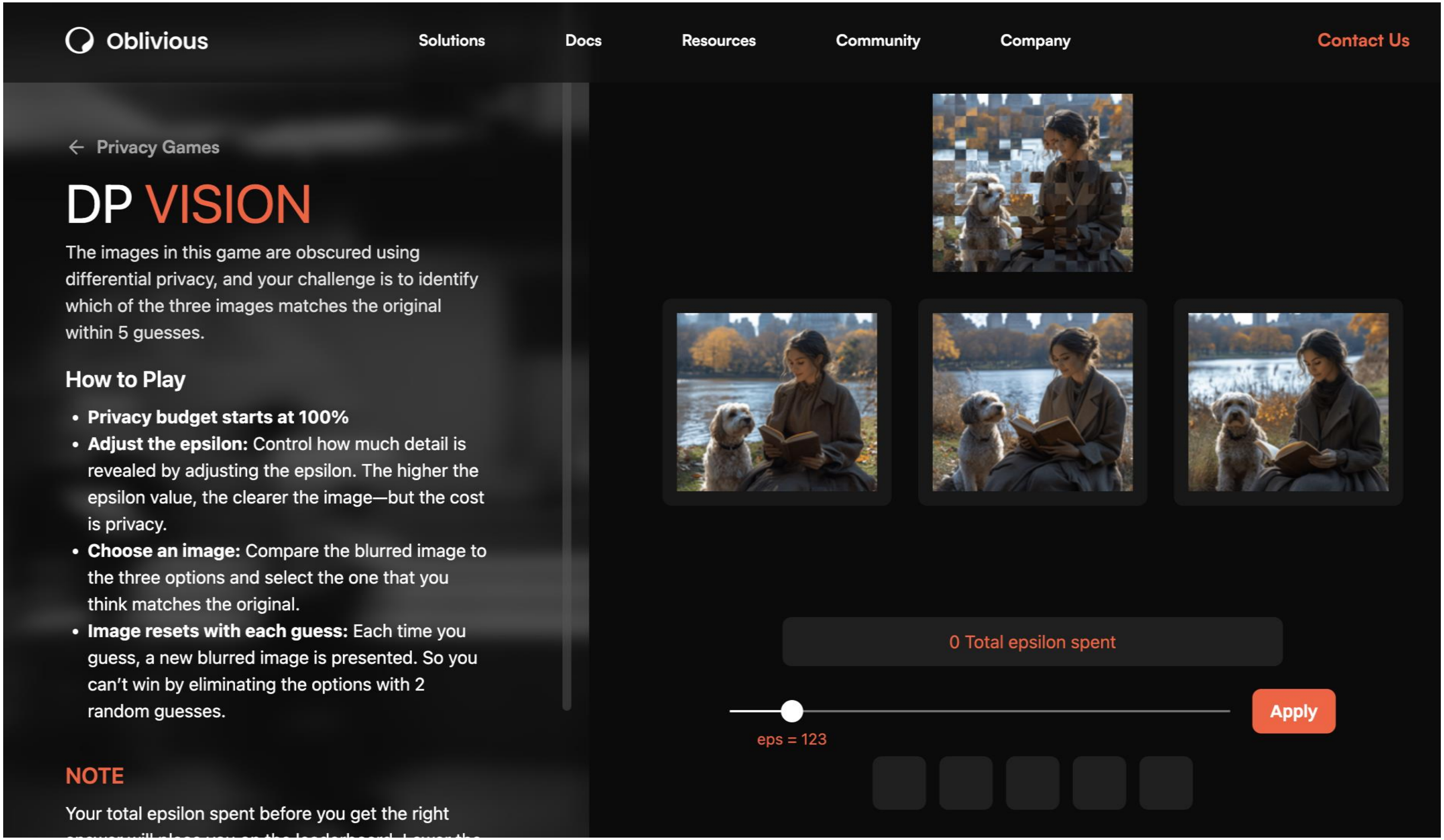
(e.g., Smart et al. 2023, Nanayakkara et al. 2023, Franzen et al. 2024, Ashena et al. 2024)

Metaphors



Karegar Alaqra Fischer-Hübner 2022

Metaphors



DP VISION

The images in this game are obscured using differential privacy, and your challenge is to identify which of the three images matches the original within 5 guesses.

How to Play

- **Privacy budget starts at 100%**
- **Adjust the epsilon:** Control how much detail is revealed by adjusting the epsilon. The higher the epsilon value, the clearer the image—but the cost is privacy.
- **Choose an image:** Compare the blurred image to the three options and select the one that you think matches the original.
- **Image resets with each guess:** Each time you guess, a new blurred image is presented. So you can't win by eliminating the options with 2 random guesses.

NOTE

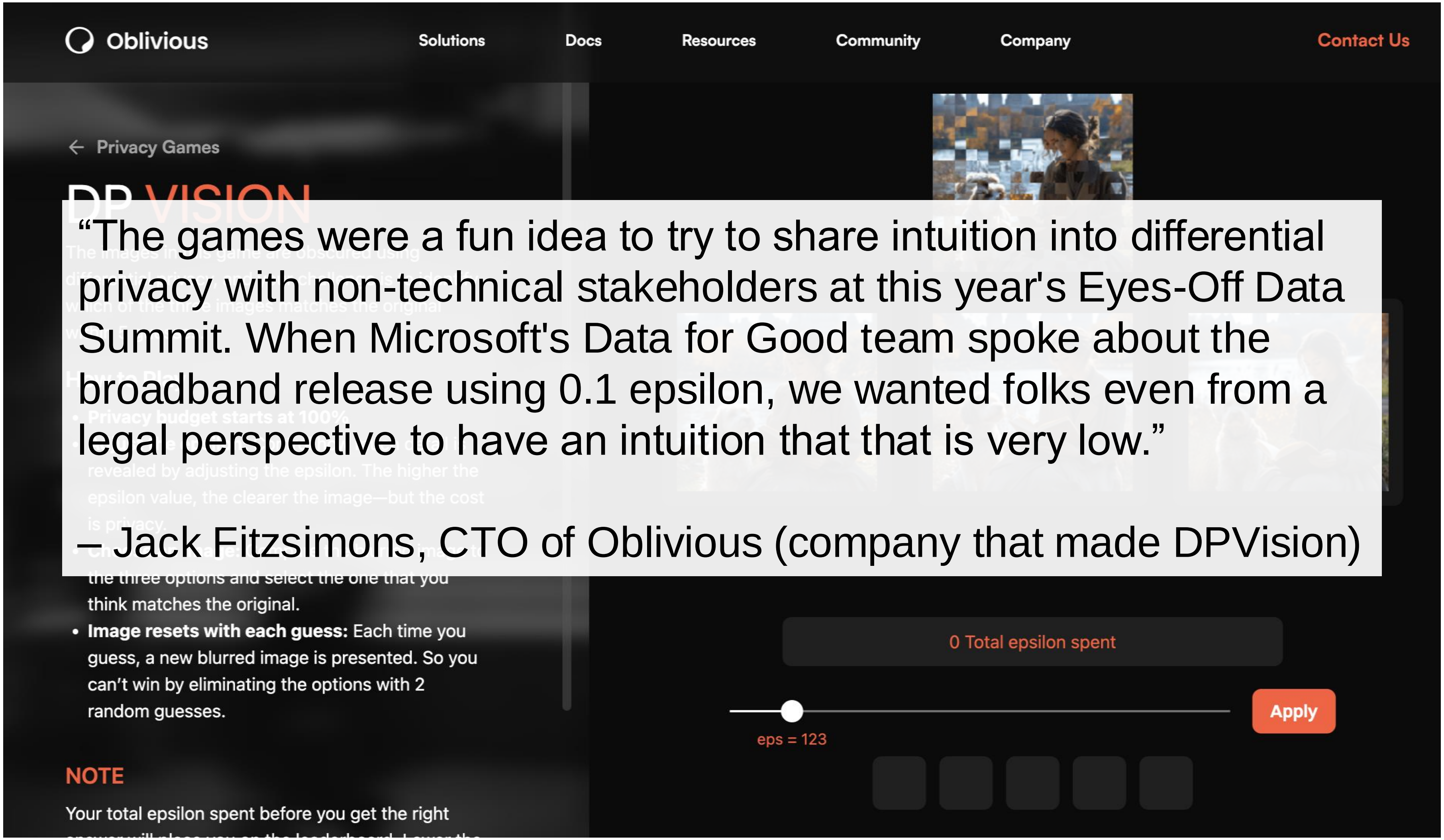
Your total epsilon spent before you get the right

0 Total epsilon spent

eps = 123

https://www.oblivious.com/games/dp-vision

Metaphors



The images in this game are obscured using differential privacy

- **Privacy budget starts at 100%**
- revealed by adjusting the epsilon. The higher the epsilon value, the clearer the image—but the cost is privacy.
- the three options and select the one that you think matches the original.
- **Image resets with each guess:** Each time you guess, a new blurred image is presented. So you can't win by eliminating the options with 2 random guesses.

NOTE

Your total epsilon spent before you get the right

> “The games were a fun idea to try to share intuition into differential privacy with non-technical stakeholders at this year's Eyes-Off Data Summit. When Microsoft's Data for Good team spoke about the broadband release using 0.1 epsilon, we wanted folks even from a legal perspective to have an intuition that that is very low.”
>
> – Jack Fitzsimons, CTO of Oblivious (company that made DPVision)

Desfontaines 2025 - https://desfontain.es/blog/dp-vision.html

Explaining epsilon: two approaches

1. ε as increase in adversary's posterior belief with data-sharing relative to without data-sharing (Wood et al. 2018)

2. ε as adversary's posterior belief with data-sharing vs. without data-sharing (Nanayakkara et al. 2023)

ε as relative increase in risk

- Communicate the maximum increase in an adversary’s posterior belief about a data subject if they share data relative to if they don’t
  - What would the adversary believe about the data subject if they see outputs of the analysis on the dataset without the data subject? How does that compare to what they’d believe about the data subject if they see outputs of the analysis on the dataset with the data subject?
- Map beliefs → financial outcome

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2018

ε as relative increase in risk

- Gertrude has a $100,000 life insurance policy. The company sets her annual premium at $1,000 = $100,000 x 0.01 (i.e., their belief of the chance she will die in the next year).

- She is considering taking part in a medical study but is concerned the study's results will lead the insurance company to believe she has a much higher chance (e.g., 0.50) of dying in the next year, resulting in a much higher annual premium (e.g., $50,000).

Can we guarantee to Gertrude that the insurance company's belief of her dying in the next year, and her insurance premium, **will not grow too much if she decides to participate vs. not participate?**

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O'Brien, Steinke, Vadhan 2018

ε as relative increase in risk

M is a randomized mechanism satisfying ε-DP

x = dataset with Gertrude; x’ = dataset without Gertrude

$d \in \{0,1\}$ = {Gertude will not die, Gertrude will die}

$$Pr[d = 1 | M(x) = y]$$

Adversary’s believes Gertrude will die next year given the analysis on the dataset with Gertrude

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2018

ε as relative increase in risk

M is a randomized mechanism satisfying ε-DP

x = dataset with Gertrude; x' = dataset without Gertrude

$d \in \{0,1\}$ = {Gertude will not die, Gertrude will die}

$$Pr[d = 1|M(x) = y]$$

$$= \frac{Pr[M(x) = y|d = 1] \cdot Pr[d = 1]}{Pr[M(x) = y]}$$

Bayes' Rule

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O'Brien, Steinke, Vadhan 2018

ε as relative increase in risk

M is a randomized mechanism satisfying ***ε***-DP

x = dataset with Gertrude; x’ = dataset without Gertrude

$d \in \{0, 1\}$ = {Gertude will not die, Gertrude will die}

$$Pr[d = 1|M(x) = y]$$

$$= \frac{Pr[M(x) = y|d = 1] \cdot Pr[d = 1]}{Pr[M(x) = y]}$$

$$\leq \frac{e^{\epsilon} Pr[M(x') = y|d = 1] \cdot Pr[d = 1]}{e^{-\epsilon} Pr[M(x') = y]}$$

Bound terms using definition of DP

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2018

ε as relative increase in risk

M is a randomized mechanism satisfying ***ε***-DP

x = dataset with Gertrude; x’ = dataset without Gertrude

$d \in \{0, 1\}$ = {Gertude will not die, Gertrude will die}

$$Pr[d = 1 | M(x) = y]$$

$$= \frac{Pr[M(x) = y | d = 1] \cdot Pr[d = 1]}{Pr[M(x) = y]}$$

$$\leq \frac{e^{\epsilon} Pr[M(x') = y | d = 1] \cdot Pr[d = 1]}{e^{-\epsilon} Pr[M(x') = y]}$$

$$\leq e^{2\epsilon} Pr[d = 1 | M(x') = y]$$

Bayes’ rule

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2018

ε as relative increase in risk

- Let’s say that the study finds a link between drinking coffee and increased risk of stroke, regardless of whether Gertrude participates. The insurance company knows Gertrude is a coffee drinker and will raise their estimate of her dying in the next year to 0.02 (resulting in a new annual premium of $2,000).

Pr[d = 1 | M(x’) = y]

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2018
See Wood, Altman, Vadhan 2020 for corrections

ε as relative increase in risk

- Let's say that the study finds a link between drinking coffee and increased risk of stroke, regardless of whether Gertrude participates. The insurance company knows Gertrude is a coffee drinker and will raise their estimate of her dying in the next year to 0.02 (resulting in a new annual premium of $2,000).

Pr[d = 1 | M(x') = y]

Pr[d = 1 | M(x) = y]

- DP guarantees that with Gertrude's participation, the insurance company's estimate will increase to at most 0.02 x $e^{2\varepsilon}$
  - For ε = 0.01, the estimate will increase to at most 0.024

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O'Brien, Steinke, Vadhan 2018
See Wood, Altman, Vadhan 2020 for corrections

$\boldsymbol{\varepsilon}$ as relative increase in risk

- Let’s say that the study finds a link between drinking coffee and increased risk of stroke, regardless of whether Gertrude participates. The insurance company knows Gertrude is a coffee drinker and will raise their estimate of her dying in the next year to 0.02 (resulting in a new annual premium of $2,000).

Pr[d = 1 | M(x’) = y]

Pr[d = 1 | M(x) = y]

- DP guarantees that with Gertrude’s participation, the insurance company’s estimate will increase to at most 0.02 x $e^{2\boldsymbol{\varepsilon}}$
  - For $\boldsymbol{\varepsilon} = 0.01$, the estimate will increase to at most 0.024

- Hence, her insurance premium will increase to at most $2,040 (i.e., her premium will increase by at most $40).

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2018
See Wood, Altman, Vadhan 2020 for corrections

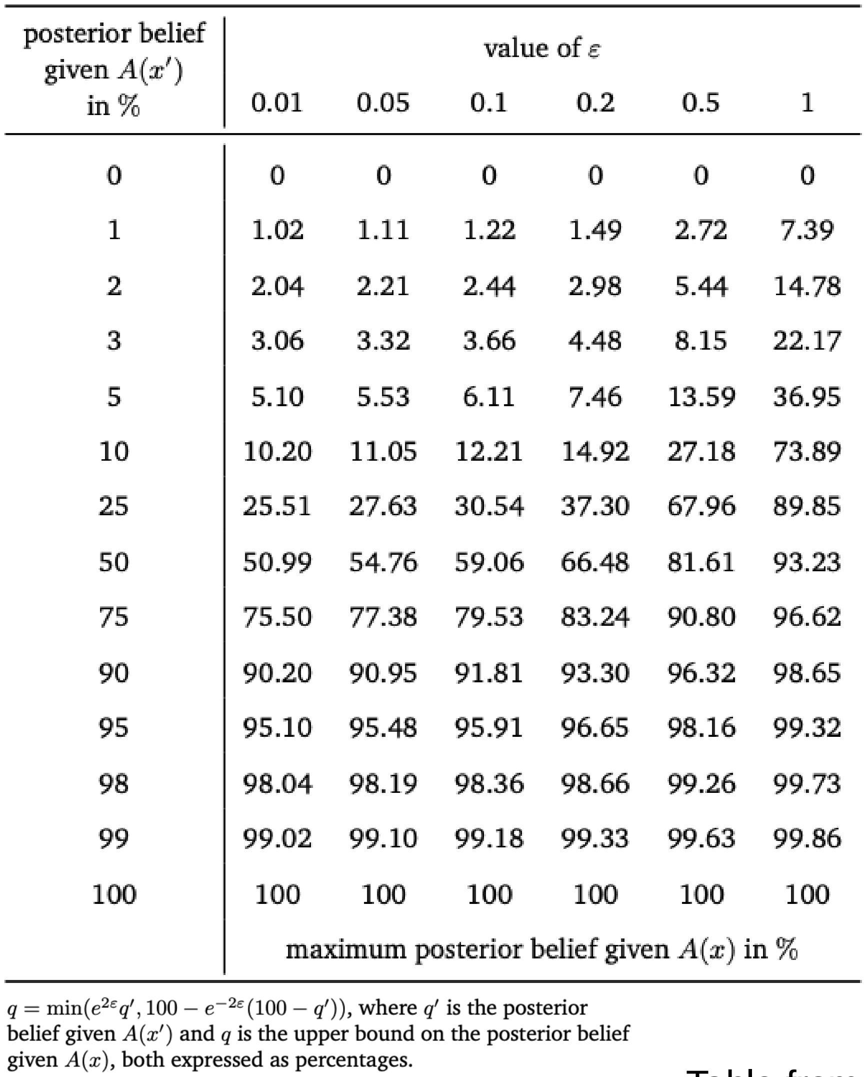
ε as relative increase in risk

- Hard to know what the insurance company will conclude about Gertrude’s death risk based on study results if Gertrude does not participate --- however, she can consider relative increases in her premium making different assumptions.

- This explanation has not been systematically evaluated with human subjects to date.
    - How might we convey several relative increases to someone without CS/stat background?

Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, O’Brien, Steinke, Vadhan 2019

ε as relative increase in risk

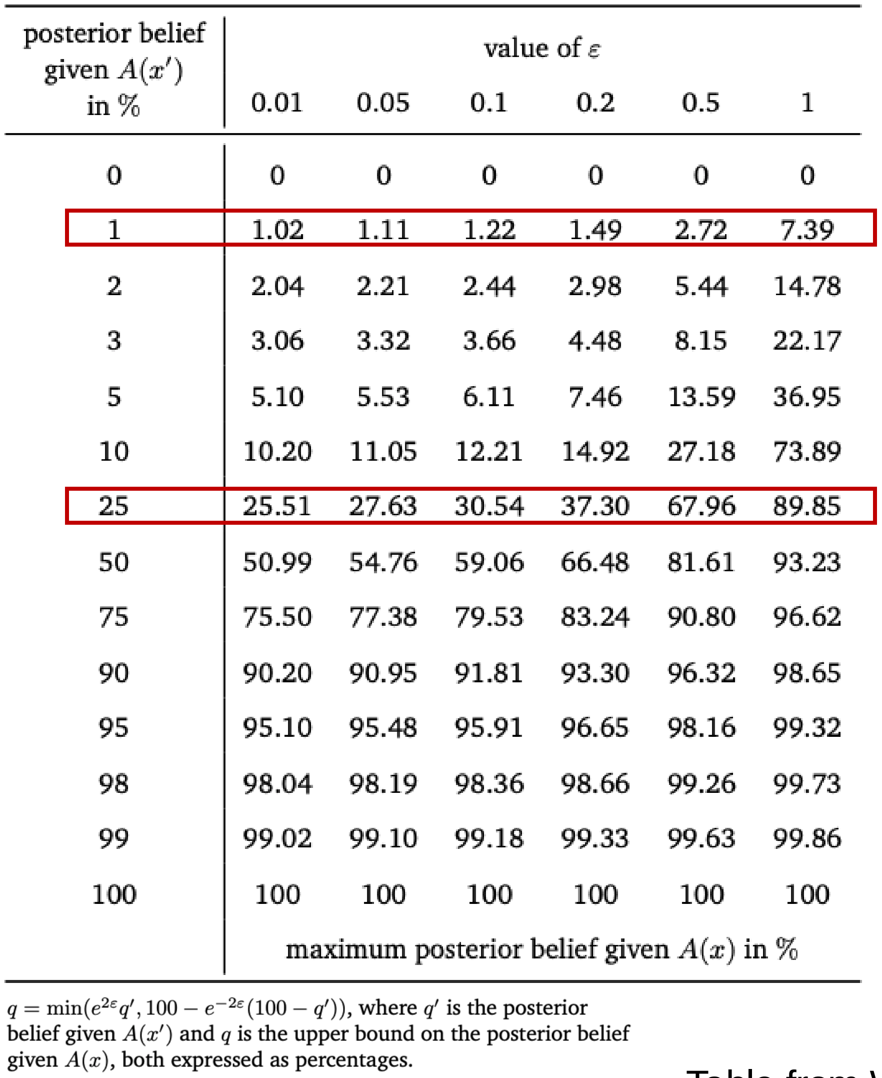


posterior belief given $A(x')$ in %	value of ε					
	0.01	0.05	0.1	0.2	0.5	1
0	0	0	0	0	0	0
1	1.02	1.11	1.22	1.49	2.72	7.39
2	2.04	2.21	2.44	2.98	5.44	14.78
3	3.06	3.32	3.66	4.48	8.15	22.17
5	5.10	5.53	6.11	7.46	13.59	36.95
10	10.20	11.05	12.21	14.92	27.18	73.89
25	25.51	27.63	30.54	37.30	67.96	89.85
50	50.99	54.76	59.06	66.48	81.61	93.23
75	75.50	77.38	79.53	83.24	90.80	96.62
90	90.20	90.95	91.81	93.30	96.32	98.65
95	95.10	95.48	95.91	96.65	98.16	99.32
98	98.04	98.19	98.36	98.66	99.26	99.73
99	99.02	99.10	99.18	99.33	99.63	99.86
100	100	100	100	100	100	100
	maximum posterior belief given $A(x)$ in %					

$q = \min(e^{2\varepsilon}q', 100 - e^{-2\varepsilon}(100 - q'))$, where q' is the posterior belief given $A(x')$ and q is the upper bound on the posterior belief given $A(x)$, both expressed as percentages.

Table from Wood, Altman, Vadhan 2020

ε as relative increase in risk



posterior belief given $A(x')$ in %	value of ε					
	0.01	0.05	0.1	0.2	0.5	1
0	0	0	0	0	0	0
1	1.02	1.11	1.22	1.49	2.72	7.39
2	2.04	2.21	2.44	2.98	5.44	14.78
3	3.06	3.32	3.66	4.48	8.15	22.17
5	5.10	5.53	6.11	7.46	13.59	36.95
10	10.20	11.05	12.21	14.92	27.18	73.89
25	25.51	27.63	30.54	37.30	67.96	89.85
50	50.99	54.76	59.06	66.48	81.61	93.23
75	75.50	77.38	79.53	83.24	90.80	96.62
90	90.20	90.95	91.81	93.30	96.32	98.65
95	95.10	95.48	95.91	96.65	98.16	99.32
98	98.04	98.19	98.36	98.66	99.26	99.73
99	99.02	99.10	99.18	99.33	99.63	99.86
100	100	100	100	100	100	100
	maximum posterior belief given $A(x)$ in %					

$q = \min(e^{2\varepsilon} q', 100 - e^{-2\varepsilon}(100 - q'))$, where q' is the posterior belief given $A(x')$ and q is the upper bound on the posterior belief given $A(x)$, both expressed as percentages.

Table from Wood, Altman, Vadhan 2020

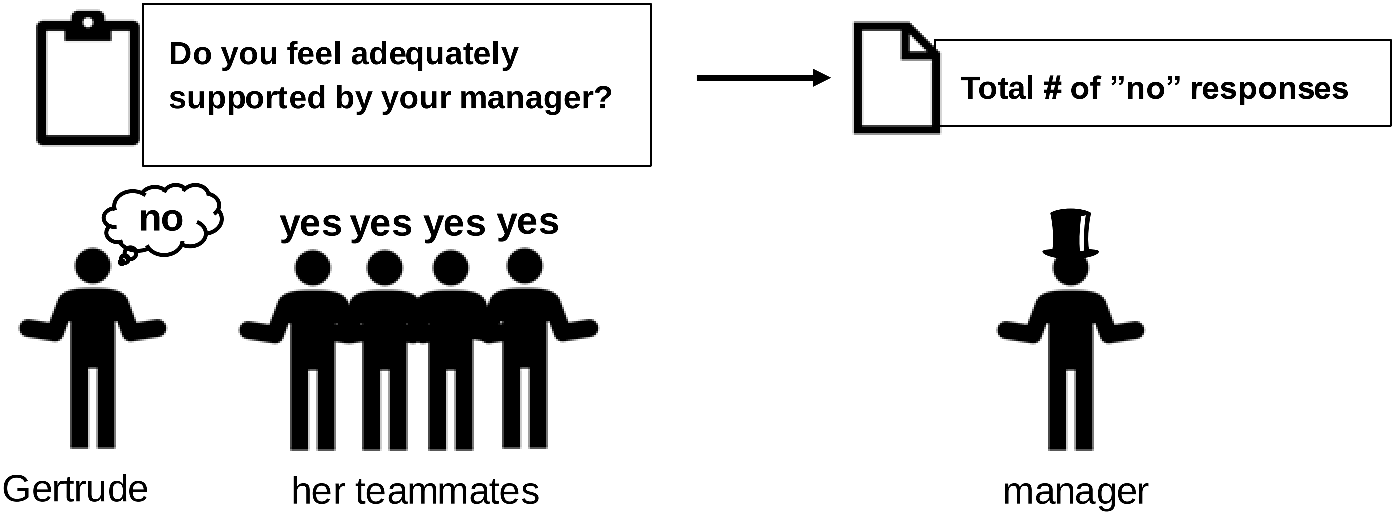
$\boldsymbol{\varepsilon}$ as absolute risk

- Communicate the adversary’s belief about the data subject if they share data vs. if they don’t share data

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Hypothetical scenario

- Now imagine Gertrude is an employee at a company that is conducting a survey. Everyone on her team must answer the question “Do you feel adequately supported by your manager?”
- Gertrude wants to respond NO, but is worried her manager will retaliate against her if he believes she responded NO.
- Her manager will be sent a report with the total number of NO responses on the survey (protected under DP). Everyone else plans to respond YES.



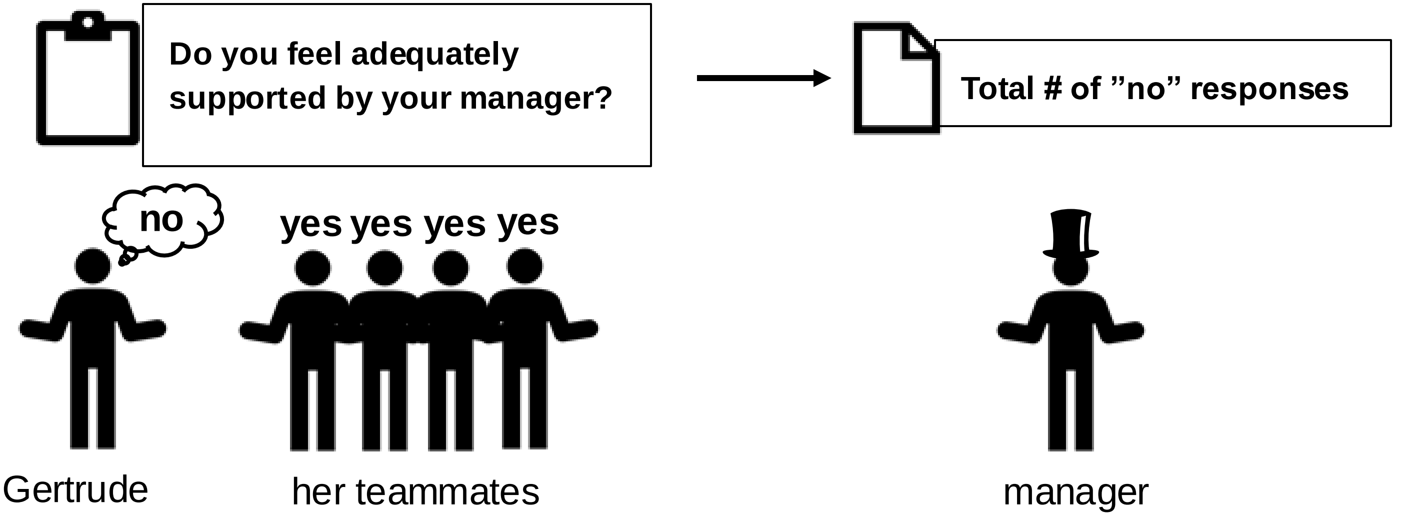
Do you feel adequately supported by your manager?

Total # of ”no” responses

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Hypothetical scenario

Can we guarantee the odds that her manager believes she responded NO will be close if she responds vs. if she doesn’t?



Do you feel adequately supported by your manager?

Total # of ”no” responses

no

yes yes yes yes

Gertrude

her teammates

manager

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Odds-based explanation

Your company will use a privacy protection method to help prevent your manager from correctly guessing anyone's response. Your company will not report exactly how many employees on your team responded NO. Instead, they will generate many potential reports by using a statistical method to modify the total number of NO responses. So, each potential report may show a number somewhat lower or higher than the actual number of NO responses. Only ONE report will be randomly sent to your manager.

If you do not participate, *x* out of 100 potential reports will lead your manager to believe you responded NO.

If you participate, *y* out of 100 potential reports will lead your manager to believe you responded NO.

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

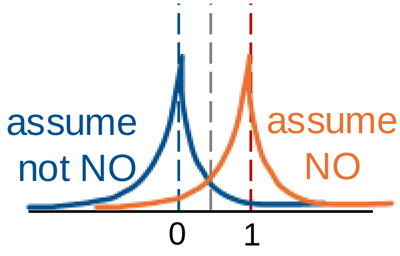
ε as absolute risk: Odds-based explanation

If you do not participate, *x* out of 100 potential reports will lead your manager to believe you responded NO.

If you participate, *y* out of 100 potential reports will lead your manager to believe you responded NO.

x and *y* depend on ε

- Model the manager as a Bayesian adversary who updates their prior belief that Gertrude responded NO.
- We will assume the manager believes Gertrude responds NO with 50% chance
- Manager determines a maximum likelihood estimate of Gertrude's response based on the differentially-private mechanism's output



Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Odds-based explanation

Probabilities reflect immediate decisions

If you **do not participate**, *x* out of 100 potential reports will lead your manager to believe you responded NO.

If you **participate**, *y* out of 100 potential reports will lead your manager to believe you responded NO.

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Odds-based explanation

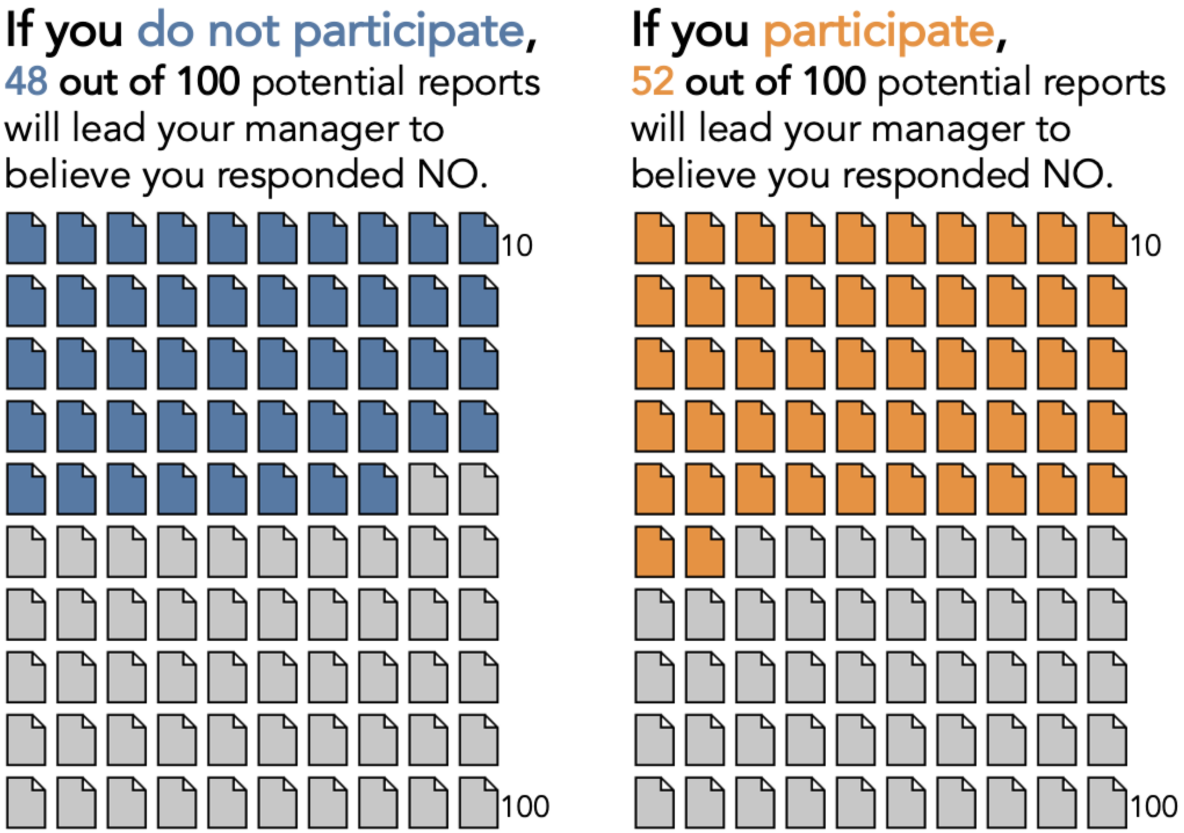
Framing probabilities as frequencies vs. percentages supports statistical reasoning (Gigerenzer & Hoffrage 95, Hoffrage & Gigerenzer 1998, Slovic 2000)

If you do not participate, ***x* out of 100** potential reports will lead your manager to believe you responded NO.

If you participate, ***y* out of 100** potential reports will lead your manager to believe you responded NO.

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Odds-based visual explanation



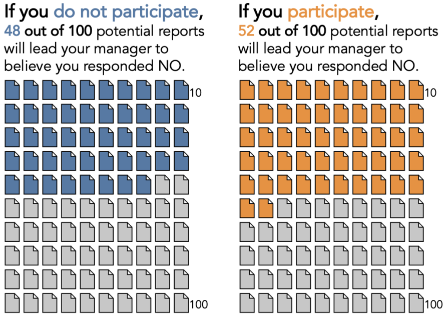
If you do not participate,
48 out of 100 potential reports will lead your manager to believe you responded NO.

If you participate,
52 out of 100 potential reports will lead your manager to believe you responded NO.

Icon arrays may improve statistical reasoning among individuals with low-numeracy skills (Galesic Garcia-Retamero Gigerenzer 2009)

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

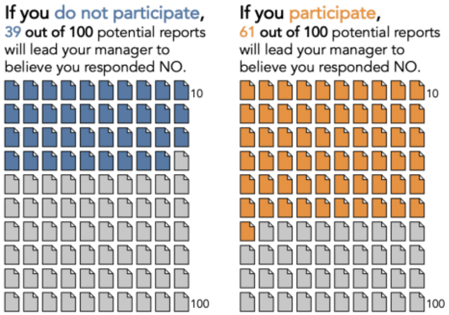
ε as absolute risk: Odds-based visual explanation



If you do not participate, 48 out of 100 potential reports will lead your manager to believe you responded NO.

If you participate, 52 out of 100 potential reports will lead your manager to believe you responded NO.

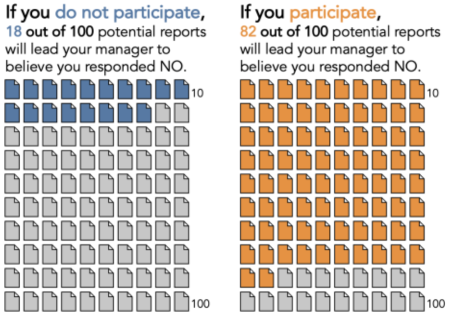
ε=0.01



If you do not participate, 39 out of 100 potential reports will lead your manager to believe you responded NO.

If you participate, 61 out of 100 potential reports will lead your manager to believe you responded NO.

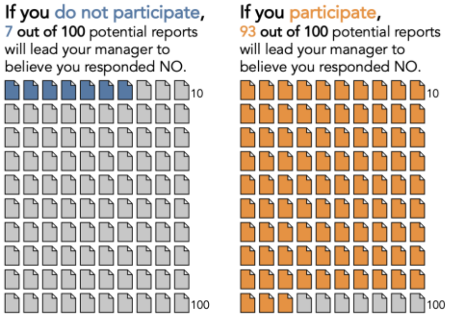
ε=0.05



If you do not participate, 18 out of 100 potential reports will lead your manager to believe you responded NO.

If you participate, 82 out of 100 potential reports will lead your manager to believe you responded NO.

ε=2



If you do not participate, 7 out of 100 potential reports will lead your manager to believe you responded NO.

If you participate, 93 out of 100 potential reports will lead your manager to believe you responded NO.

ε=4

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Evaluation

- objective risk comprehension
- subjective privacy understanding
- self-efficacy
  - Confidence deciding
  - Feelings of having enough information
- willingness to share data

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Evaluation

- Between-subjects vignette survey study

- Hypothetical workplace scenario with a data-sharing decision + the explanation instantiated under one of four ε (0.1, 0.5, 2, 4) or a control

- Half were told to imagine they can participate or opt-out, the other half were told to imagine they had to respond but could lie

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Evaluation

- Controls
  - Privacy control based on prior work (Xiong Wang Li Jha 2020) (mix between “enables” and techniques” from before; no mention of ε)
  - No-privacy, deterministic control

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Results

- Participants were over 2x as likely to answer an additional objective risk comprehension question correctly with the odds-based visual explanation vs. the no-privacy, deterministic control.

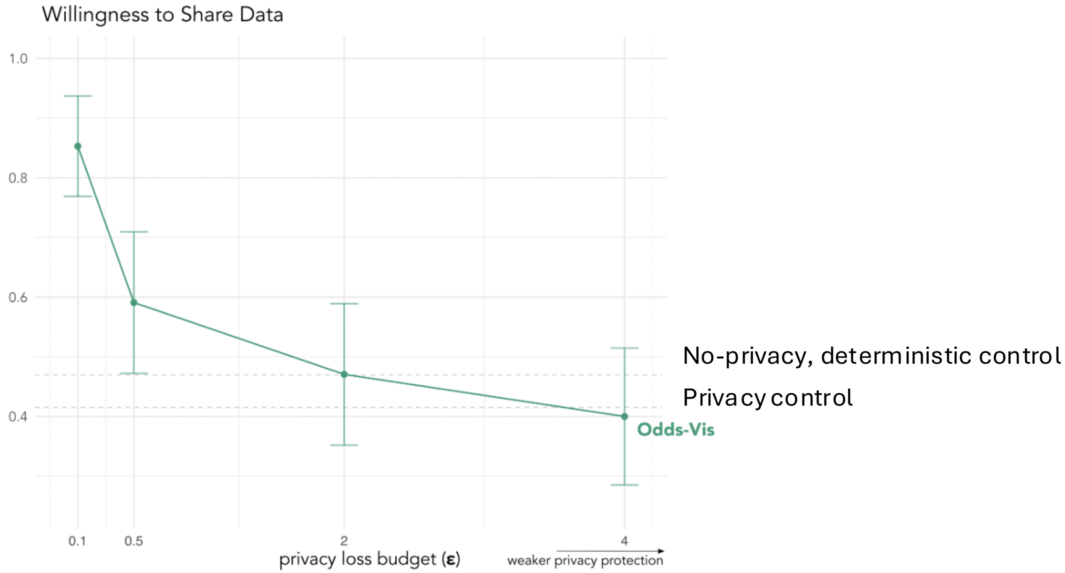
- Compared to the privacy control, the odds-based visual explanation improved self-efficacy (feelings of having enough information to decide) (O.R. = 1.7).

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Results

Compared to the privacy control from prior work, participants were nearly 2x as likely to share data when given the odds-based visual explanation

Participants were more likely to share data with smaller ε



Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

ε as absolute risk: Qualitative feedback

- **Utility of responses**
  - “The random process completely obfuscates the true number of NO responses; that is great for employee anonymity, but is kind of useless for the manager.”
- **Scenario context**
  - “Who else gets the survey results? Does HR get the correct information and so I trust that HR will help me if the boss retaliates.”
- **Honesty**
  - “I’ve actually been in a somewhat similar situation, and I was retaliated against. Still, despite that experience, I think it is important to be honest about what is going on.”

Nanayakkara Smart Cummings Kaptchuk Redmiles 2023

Explaining epsilon: two approaches

1. ε as increase in adversary's posterior belief with data-sharing relative to without data-sharing (Wood et al. 2018)

2. ε as adversary's posterior belief with data-sharing vs. without data-sharing (Nanayakkara et al. 2023)

Similarities:

- Both make some comparison about what might happen with sharing data vs. not sharing data, aligning with the definition of DP (as opposed to, with DP versus without DP)
- Both require making some assumptions (e.g., about the adversary's attack procedure or what they will learn on the analysis without the data subject's info)
- Both express a worst-case guarantee

Differences:

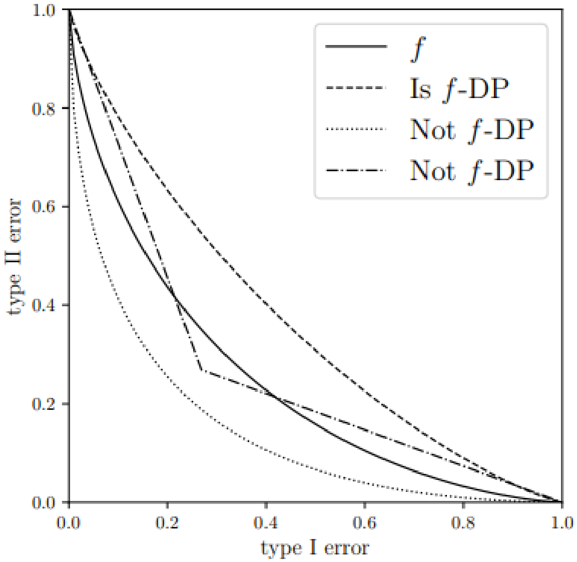
- Relative vs. absolute probabilities
- Presentation (text vs. visual; frequency- vs. percentage-framing of probabilities)

Takeaways

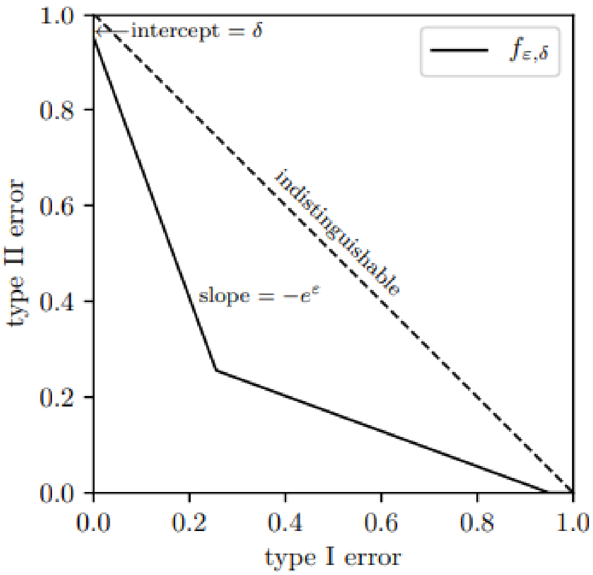
- Currently-used explanations of DP don’t always set accurate privacy expectations
- Explanations of DP include metaphors, visualizations, and text descriptions – which approach you use depends on context and needs of data subjects
- There are multiple approaches to numerically explaining DP’s guarantees. They tend to involve interpreting the definition in ways that align with a data subject’s decision (share data vs don’t share data)
- Best practices around explaining DP are still evolving – could be a topic of research for your project! Consider not just data subjects, but also other parties

f-DP

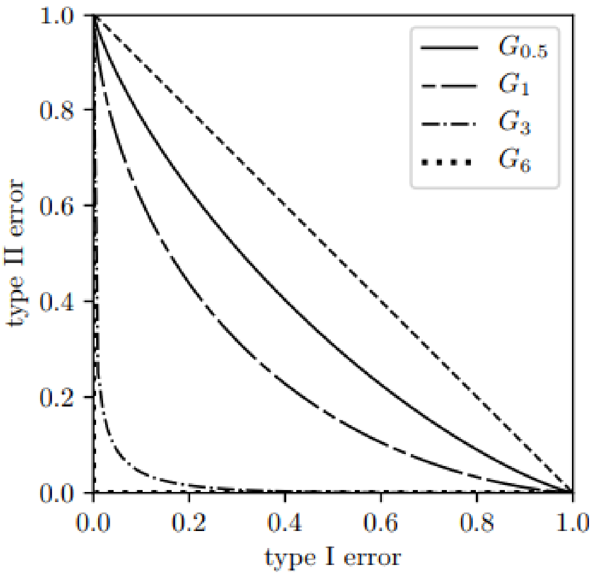
Privacy guarantees specified by a $f : [0,1] \to [0,1]$
s.t. $\mathrm{FNR} \geq f(\mathrm{FPR})$ at all $\mathrm{FPR} \in [0,1]$
in distinguishing $H_0 = M(x)$ from $H_1 = M(x')$ for
$x \sim x'$



Illustrating the def



(ε, δ)-DP as f-DP



Gaussian mechanism

f-DP is equivalent to giving a full ε vs. δ curve (rather than a single pair).