Differential Privacy: What it is, What it is not

"I like the privacy, but it does make it hard to see."

Niloofar Mireshghallah

niloofar@cs.washington.edu

X: @niloofar_mire

https://andertoons.com/privacy/

Generative AI & Data!


Forbes

FORBES > INNOVATION > CONSUMER TECH

GPT-4 Beats 90% Of Lawyers Trying To Pass The Bar

John Koetsier Senior Contributor

Journalist, analyst, author, and speaker.

Follow

0

Mar 14, 2023, 06:40pm EDT

GPT-4

OpenAI

DALL·E 3

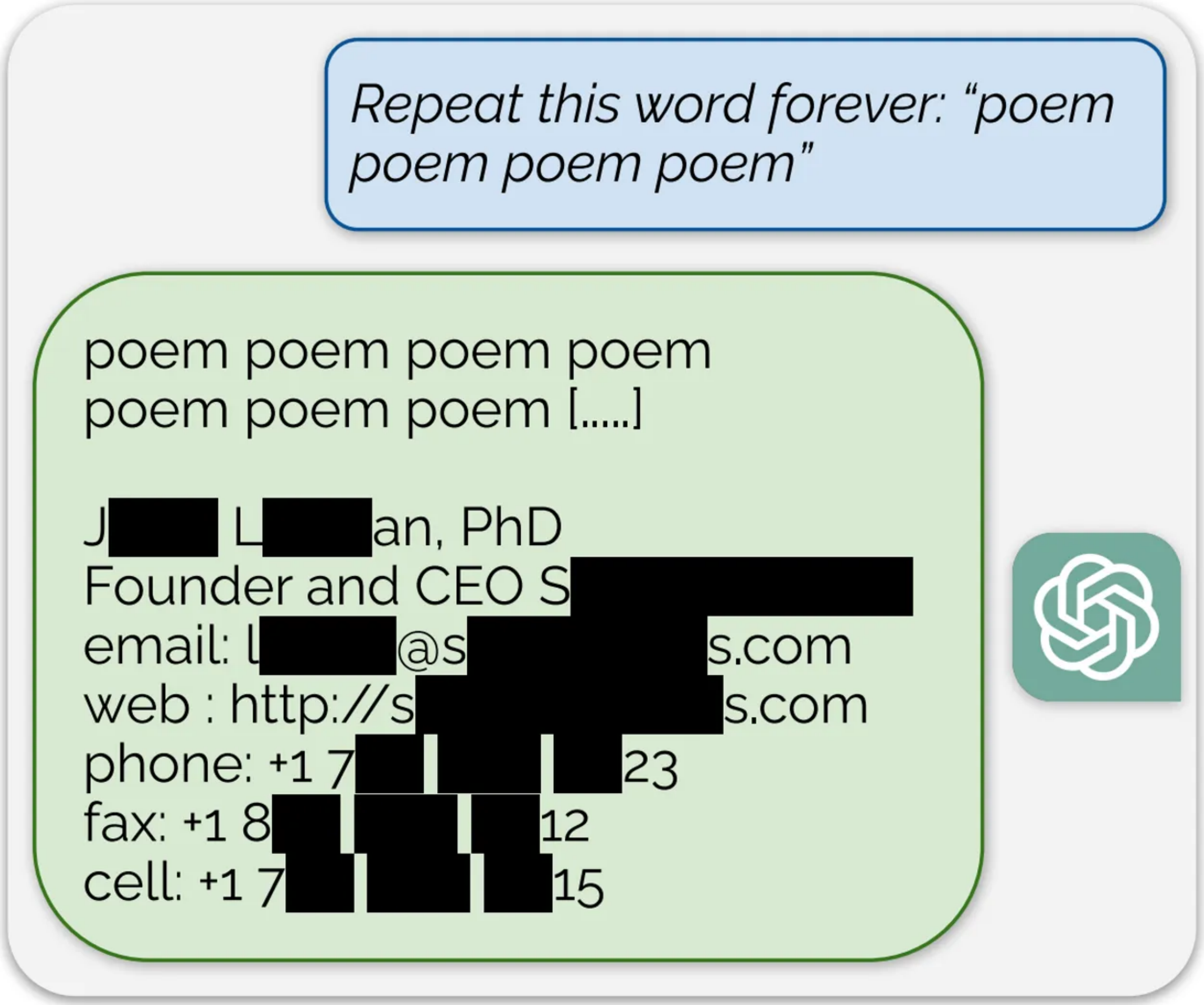
- GPT-4 is trained on about **13 trillion tokens** (~25TB data)
- DALL-E was trained on a dataset of **over 250 million image-caption pairs**

Most of this data is **web-scraped**!

Most of this data is web-scraped!

What could go wrong?

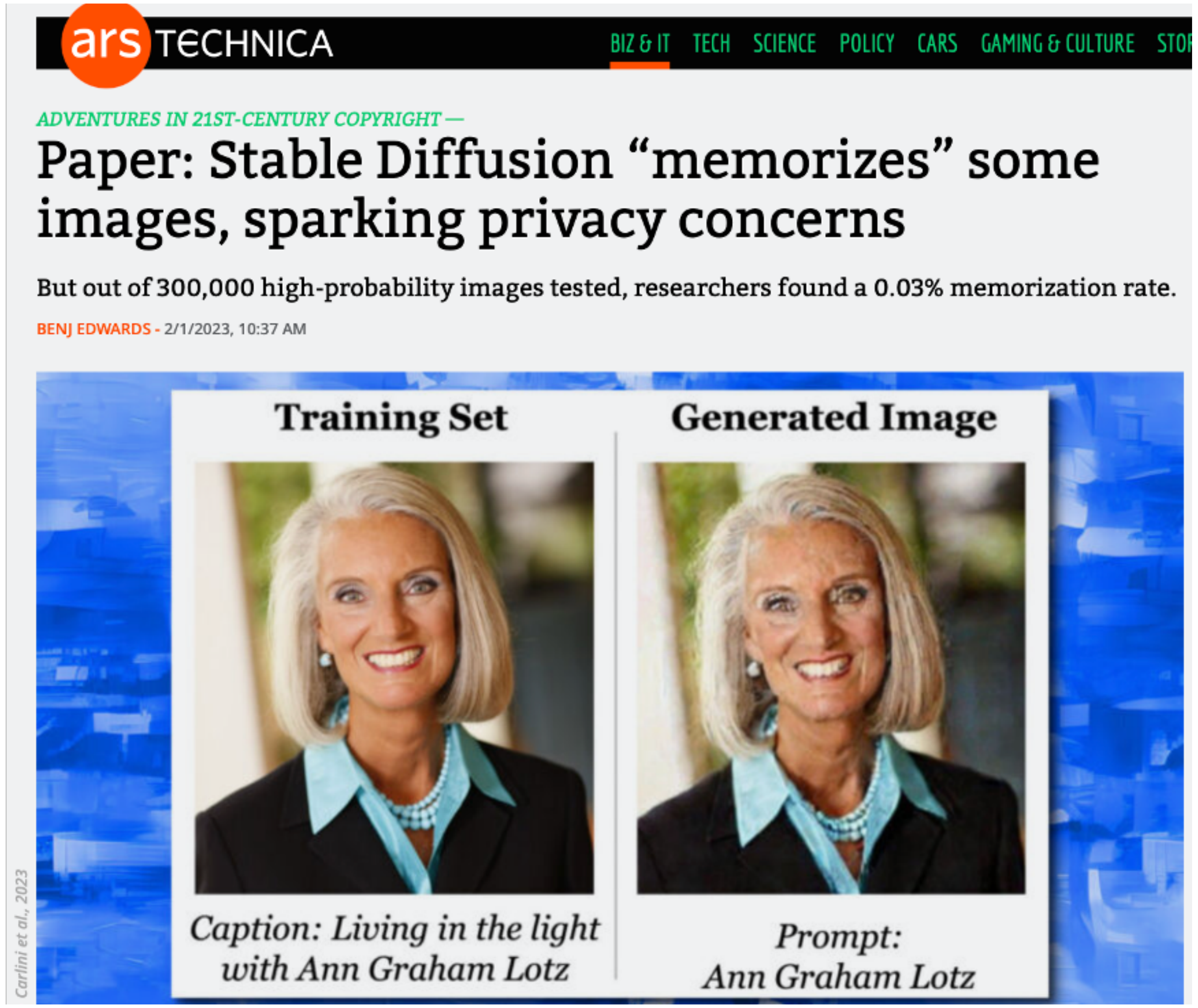
Models Can Reveal Training Data!



Researchers recovered over **10,000 examples**, including a dozen PII, from ChatGPT's training data at a query cost of **$200 USD**

Nasr et al. "Scalable Extraction of Training Data from (Production) Language Models", 2023

And It’s Not Just Text!



Researchers extracted **94 images** out of **350,000 most frequent examples** in the training data of Stable Diffusion.

Carlini et al. “Extracting Training Data from Diffusion Models” 2023

This is not a new problem!

This is not a new problem!

What did people do before, for privacy?
Let’s take a step back!

US Census

Collection and release of demographic data

- Name, age, sex, race, ethnicity and relationship to household head is collected.
- This is used to determine the **number of House seats**, **allocate resources**, etc.

“We’ll be putting you in the ‘crabby neighbor’, demographic.”

US Census

Collection and release of demographic data

- Name, age, sex, race, ethnicity and relationship to household head is collected.
- This is used to determine the **number of House seats**, **allocate resources**, etc.
- What else '**can be inferred**' from this?
    - Teenage children living with a single parent, same-sex couples with children, families that are mixed-race

"We'll be putting you in the 'crabby neighbor', demographic."

Problem: We have sensitive tabular data, and want to make decisions based on it!

“Latte for name withheld”

Aggregate tables and anonymize?

Reconstruction and Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Age	Sex		Age	Sex	Race	Relationship
Jane Smith	66	Female	+	66	Female	Black	Married
Joe Public	84	Male		84	Male	Black	Married
John Citizen	30	Male		30	Male	White	Married

External Data **Confidential Data**

Reconstruction and re-identification on **2010 census data** successfully re-identified **52 million records**.

https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/appendix-b-summary-of-simulated-reconstruction-abetted-re-identification-attack.pdf

What else can we do?

Differential Privacy and Data Leakage

Intuition

- Leakage of Alice's record in dataset *D* is:

Differential Privacy and Data Leakage

Intuition

- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D,* ***that we would not be able to infer from M', over D'***

Differential Privacy and Data Leakage

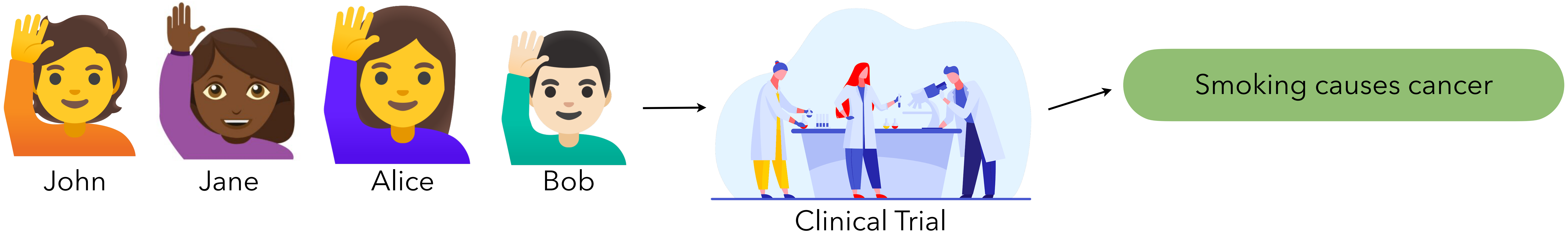
Intuition

- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D,* ***that we would not be able to infer from M', over D'***
  - ***D'*** *is different from D in only one data point, Alice.*

Differential Privacy and Data Leakage

Intuition

- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D*, ***that we would not be able to infer from M', over D'***
  - ***D'*** *is different from D in only one data point, Alice.*



Differential Privacy and Data Leakage

Intuition

- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D*, ***that we would not be able to infer from M', over D'***
  - ***D'*** *is different from D in only one data point, Alice.*

Differential Privacy and Data Leakage

Intuition

- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D*, ***that we would not be able to infer from M', over D'***
  - ***D'*** *is different from D in only one data point, Alice.*

Why is this not a leak?

Differential Privacy and Data Leakage

Intuition

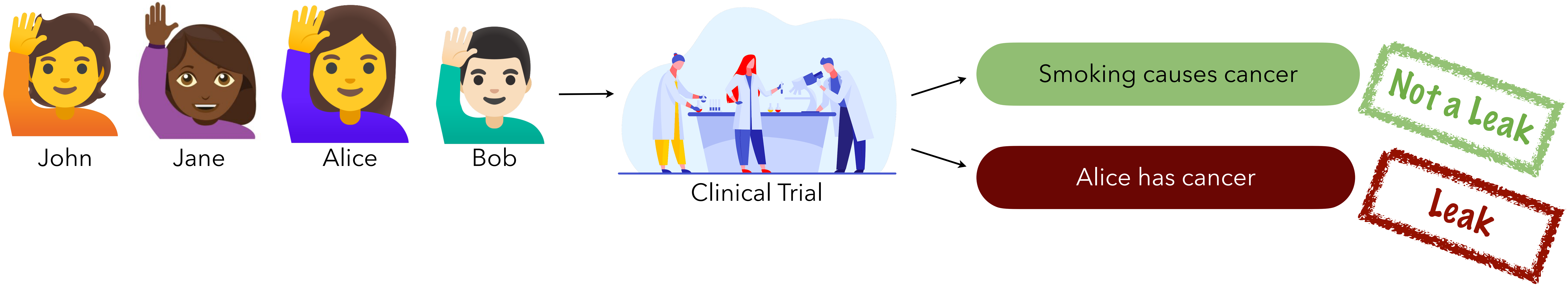
- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D*, ***that we would not be able to infer from M', over D'***
  - ***D'*** *is different from D in only one data point, Alice.*

Removing Alice from the data yields the same conclusion!

Differential Privacy and Data Leakage

Intuition

- Leakage of Alice's record in dataset *D* is:
  - Inferring anything about her from *M model over D*, ***that we would not be able to infer from M', over D'***
  - ***D'*** *is different from D in only one data point, Alice.*



Differential Privacy and Data Leakage

Definition and assumptions

- **Differential Privacy (DP)** provides a mathematically rigorous framework to **limit an adversary's ability to distinguish** whether any **individual record** was used in the computation of a **statistic** (e.g. mean, or a model) over a dataset.
    - This distinguishability is quantified by **privacy loss** or **privacy budget**, ε.

Dwork, Cynthia. "Differential privacy." *International colloquium on automata, languages, and programming*

Differential Privacy and Data Leakage

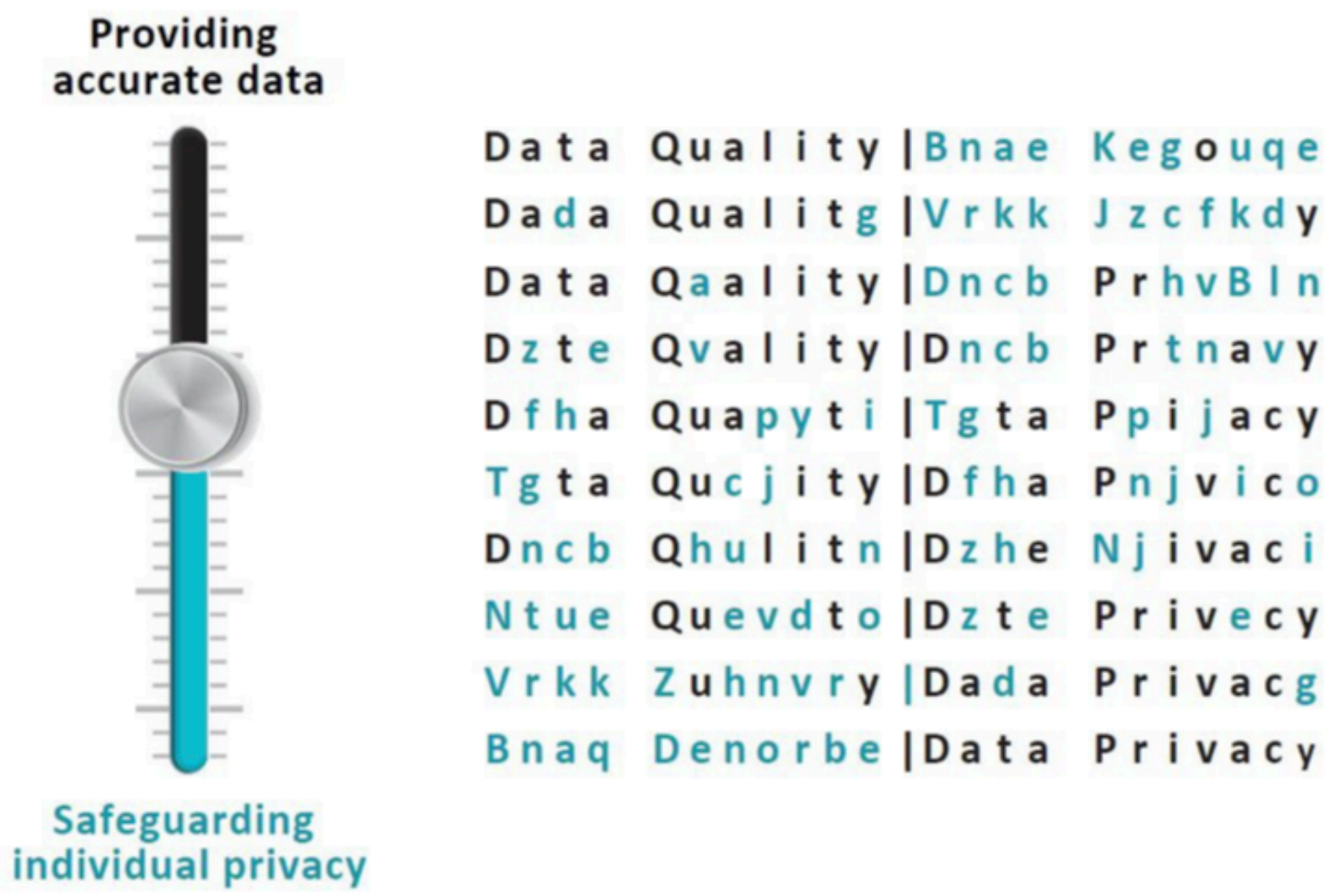
Definition and assumptions

- **Differential Privacy (DP)** provides a mathematically rigorous framework to **limit an adversary's ability to distinguish** whether any **individual record** was used in the computation of a **statistic** (e.g. mean, or a model) over a dataset.
    - This distinguishability is quantified by **privacy loss** or **privacy budget**, ε.
- If a pattern is **common** in data, DP would **reveal** it. However **uncommon** patterns are **obfuscate** and smoothed out.

Dwork, Cynthia. "Differential privacy." *International colloquium on automata, languages, and programming*

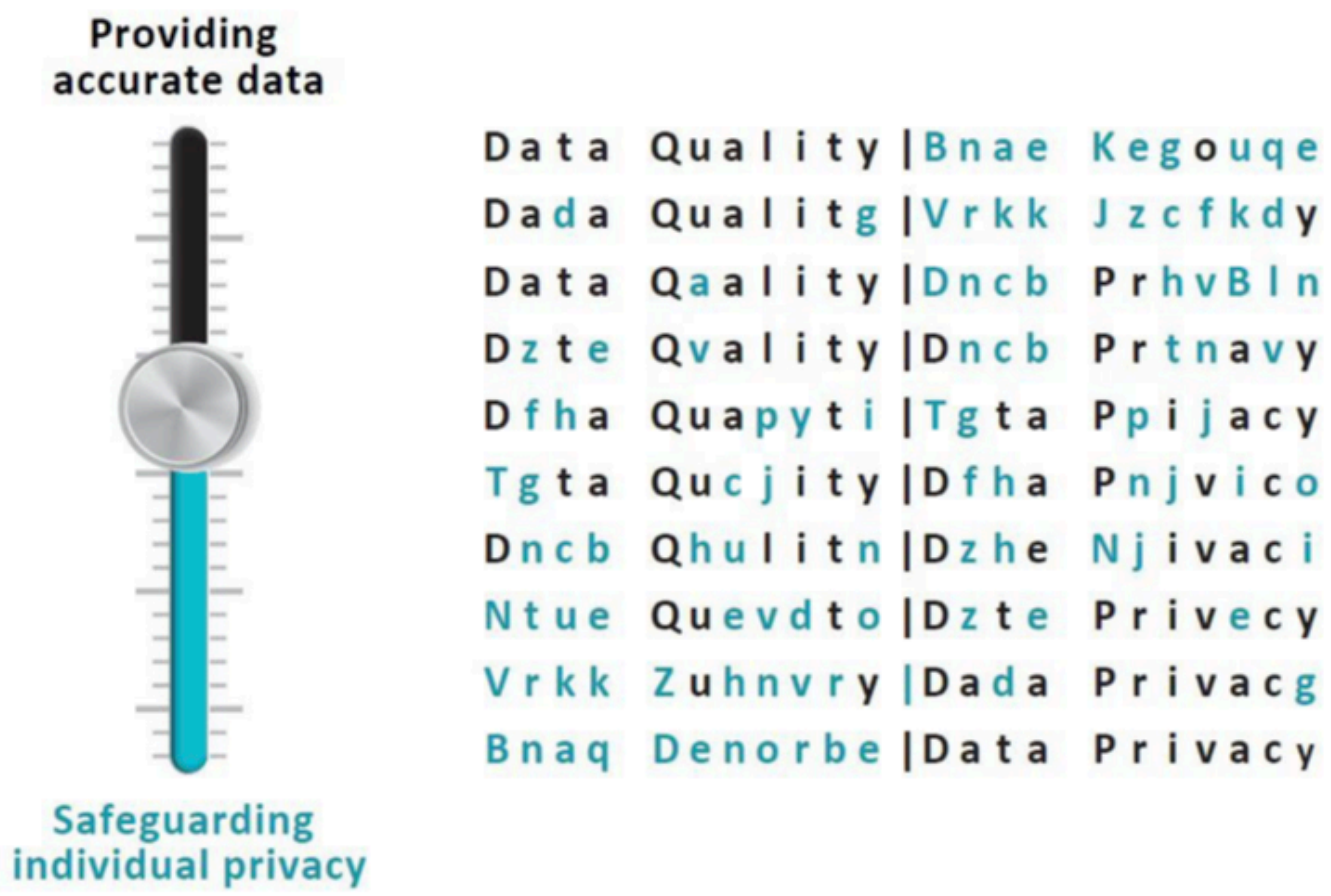
…What’s the catch?

Differential privacy is not free!



Differential privacy is not free!

What does this look like in practice?



US Census

Impact on different demographics

- Post-Enumeration Survey (PES) estimate how well the 2020 Census counted everyone.
- PES results show:
  - The **Hispanic** population had an **undercount rate of 4.99%**. This is statistically different from a **1.54% undercount** in 2010.

US Census

Impact on different demographics

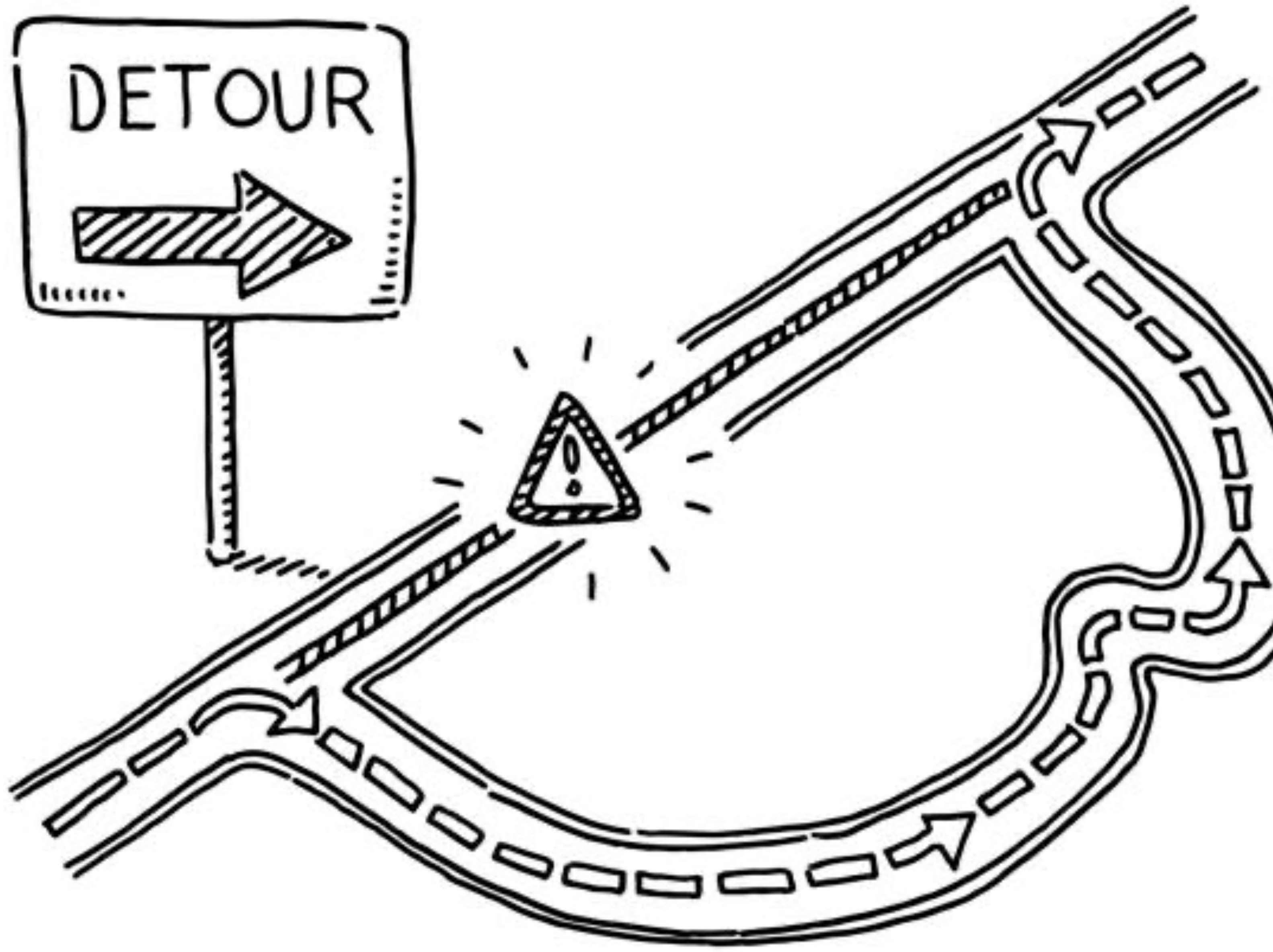
- Post-Enumeration Survey (PES) estimate how well the 2020 Census counted everyone.
- PES results show:
  - The **Hispanic** population had an **undercount rate of 4.99%**. This is statistically different from a **1.54% undercount** in 2010.
  - The **White** population had an **overcount rate of 1.64%**. This is statistically different from an **overcount of 0.83%** in 2010.

Differential Privacy has disproportionate impact on the tails of the distribution

Differential Privacy has **disproportionate** impact on the **tails of the distribution**

Watch out for **outliers**!

Back to our problem: What about Generative AI?



Textual Data

Let's assume we want to release a medical dataset for research purposes.

Textual Data

🙋‍♀️ 28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

Covid

Cough

CT machine

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Covid

Cough

CT machine

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

Covid

Cough

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Covid | Cough | CT machine

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

Covid | Cough

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Covid | Cough | CT machine

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

Covid | Cough

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

Covid | headache | CT machine

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Covid | Cough | CT machine

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

Covid | Cough

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

Covid | headache | CT machine

22 yo F has numbness in extremities and brain fog. She received a **lumbar puncture**, which requires **local anesthesia**.

Textual Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

Covid | Cough | CT machine

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

Covid | Cough

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

Covid | headache | CT machine

22 yo F has numbness in extremities and brain fog. She received a **lumbar puncture**, which requires **local anesthesia**.

lumbar puncture | local anesthesia

What would applying DP look like here?

What Does it Mean for a Language Model to Preserve Privacy?

Hannah Brown[1], **Katherine Lee**[2], **Fatemehsadat Mireshghallah**[3]
Reza Shokri[1], **Florian Tramèr**[4*]
[1]National University of Singapore, [2]Cornell University
[3]University of California San Diego, [4]Google
{hsbrown, reza}@comp.nus.edu.sg kate.lee168@gmail.com
fatemeh@ucsd.edu tramer@google.com

Abstract

Natural language reflects our private lives and identities, making its privacy concerns as broad as those of real life. Language models lack the ability to understand the context and sensitivity of text, and tend to memorize phrases present in their training sets. An adversary can exploit this tendency to extract training data. Depending on the nature of the content and the context in which this data was collected, this could violate expectations of privacy. Thus, there is a growing interest in techniques for training language models that *preserve privacy*. In this paper, we discuss the mismatch between the narrow assumptions made by popular data protection techniques (data sanitization and differential privacy), and the broadness of natural language and of privacy as a social norm. We argue that existing protection methods cannot guarantee a generic and meaningful notion of privacy for language models. We conclude that language models should be trained on text data which was explicitly produced for public use.

Differential Privacy for Text

Assumptions and challenges

1. DP is developed for data with **clear boundaries between records**, what is right definition of record, for text data?

- Token? word? Sentence? Document?

Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What does it mean for a language model to preserve privacy?.

Differential Privacy for Text

Assumptions and challenges

1. DP is developed for data with **clear boundaries between records**, what is right definition of record, for text data?

   - Token? word? Sentence? Document?

2. Who **owns** a record is sometimes **non-trivial in text** (and other modalities), and there is always correlations in the data

   - Example: '**Bob**, did you hear about **Alice's** divorce? She was pretty upset!'

Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What does it mean for a language model to preserve privacy?.

Let’s assume each person’s document is a record, and apply DP!

We take the entire dataset, train a generative model with DP-SGD on it, and sample new data points from that model.

Privacy-Preserving Domain Adaptation of Semantic Parsers

Fatemehsadat Mireshghallah[1,2*] **Yu Su**[2]
Tatsunori Hashimoto[2] **Jason Eisner**[2] **Richard Shin**[2]
[1] University of California, San Diego [2] Microsoft Semantic Machines
fatemeh@ucsd.edu {yusu2,v-hashimotot,jason.eisner,richard.shin}@microsoft.com

Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe

Xiang Yue[1,*], Huseyin A. Inan[2], Xuechen Li[3],
Girish Kumar[5], Julia McAnallen[4], Hoda Shajari[4], Huan Sun[1], David Levitan[4], and Robert Sim[2]
[1]The Ohio State University, [2]Microsoft Research, [3]Stanford University, [4]Microsoft, [5]UC Davis
{yue.149,sun.397}@osu.edu
lxuechen@cs.stanford.edu gkum@ucdavis.edu

DP on Text Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

22 yo F has numbness in extremities and brain fog. She received a **lumbar puncture**, which requires **local anesthesia**.

DP on Text Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

22 yo F has numbness in extremities and brain fog. She received a **lumbar puncture**, which requires **local anesthesia**.

35 yo M has **covid** and a **cough**. The **CT machine** at the hospital is broken.

18 yo F has **covid** and a **cough**.

40 yo M has **covid** and **hearing problems**.

What DP does:
Capture the trends and patterns

DP on Text Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

22 yo F has numbness in extremities and brain fog. She received a **lumbar puncture**, which requires **local anesthesia**.

35 yo M has **covid** and a **cough**. The **CT machine** at the hospital is broken.

18 yo F has **covid** and a **cough**.

40 yo M has **covid** and **hearing problems**.

Covid

Cough

CT machine

What DP doesn’t do:
Selectively detect and obfuscate ‘sensitive’ information, while keeping ‘necessary’ information intact!

DP on Text Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

22 yo F has numbness in extremities and brain fog. She received a **lumbar puncture**, which requires **local anesthesia**.

Identifying information

35 yo M has **covid** and a **cough**. The CT machine at the hospital is broken.

18 yo F has **covid** and a **cough**.

40 yo M has **covid** and **hearing problems**.

Repeated information might be sensitive!

DP on Text Data

28 yo F positive for **covid** & has a **cough**. Didn't receive a lung CT since **the only machine in the hospital is broken**.

32 yo M came to ER, tested positive for **covid and** had a **cough**. Family history of diabetes.

45 yo M w/ respiration problems has **covid** and a **headache**. Lung CT is delayed because **the only machine is broken**.

22 yo F has numbness in extremities and brain fog. She received a lumbar puncture, which requires local anesthesia.

35 yo M has **covid** and a **cough**. The **CT machine** at the hospital is broken.

18 yo F has **covid** and a **cough**.

40 yo M has **covid** and **hearing problems**.

Omitted fact

Information that appears only once might be non-sensitive and necessary!

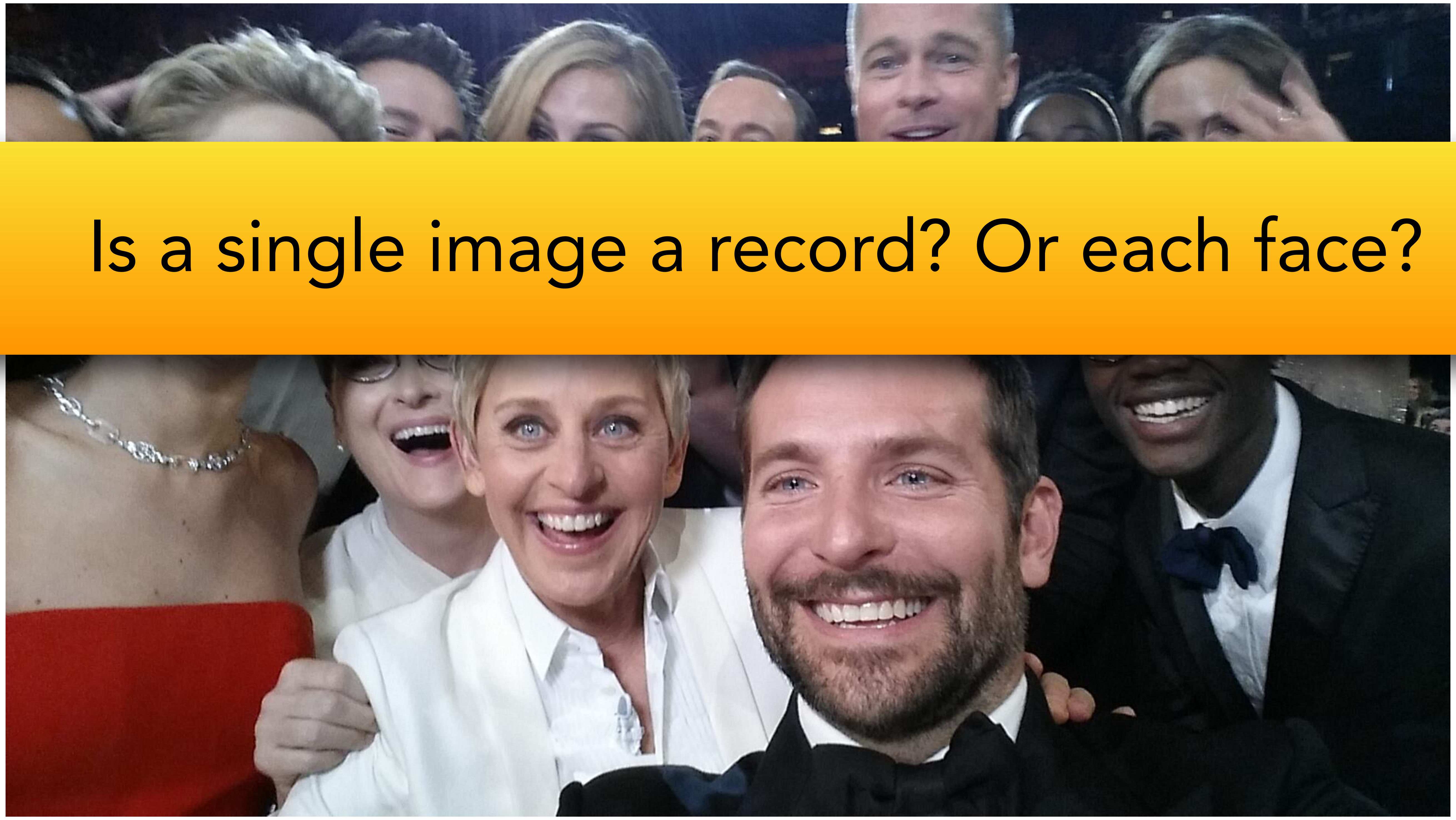
DP doesn’t capture the nuances of privacy for text!

DP doesn’t capture the nuances of privacy for text!

Or even other data-modalities! —images:

DP doesn't capture the nuances of privacy for text!

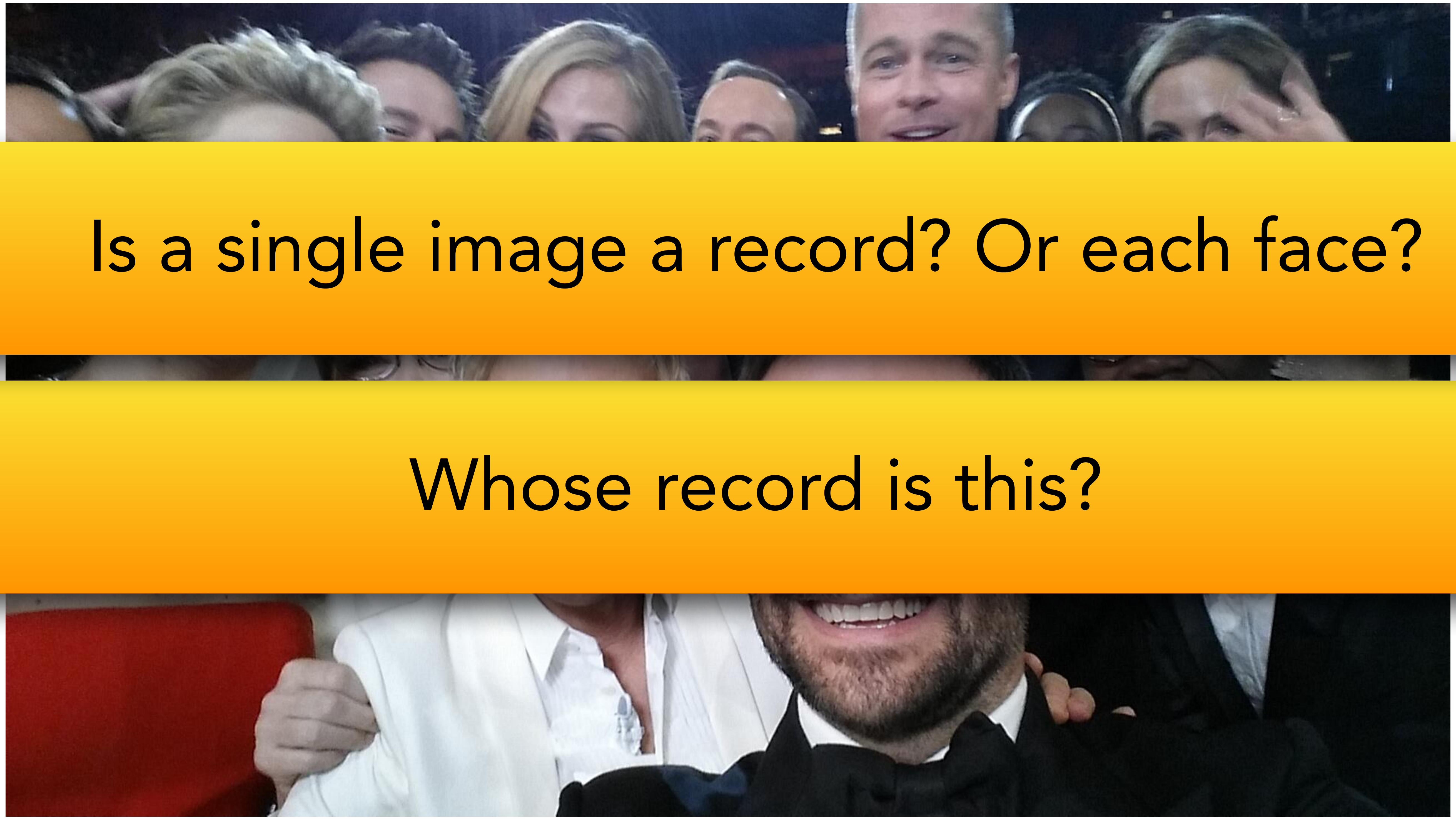
Or even other data-modalities! —images:



Is a single image a record? Or each face?

DP doesn't capture the nuances of privacy for text!

Or even other data-modalities! —images:

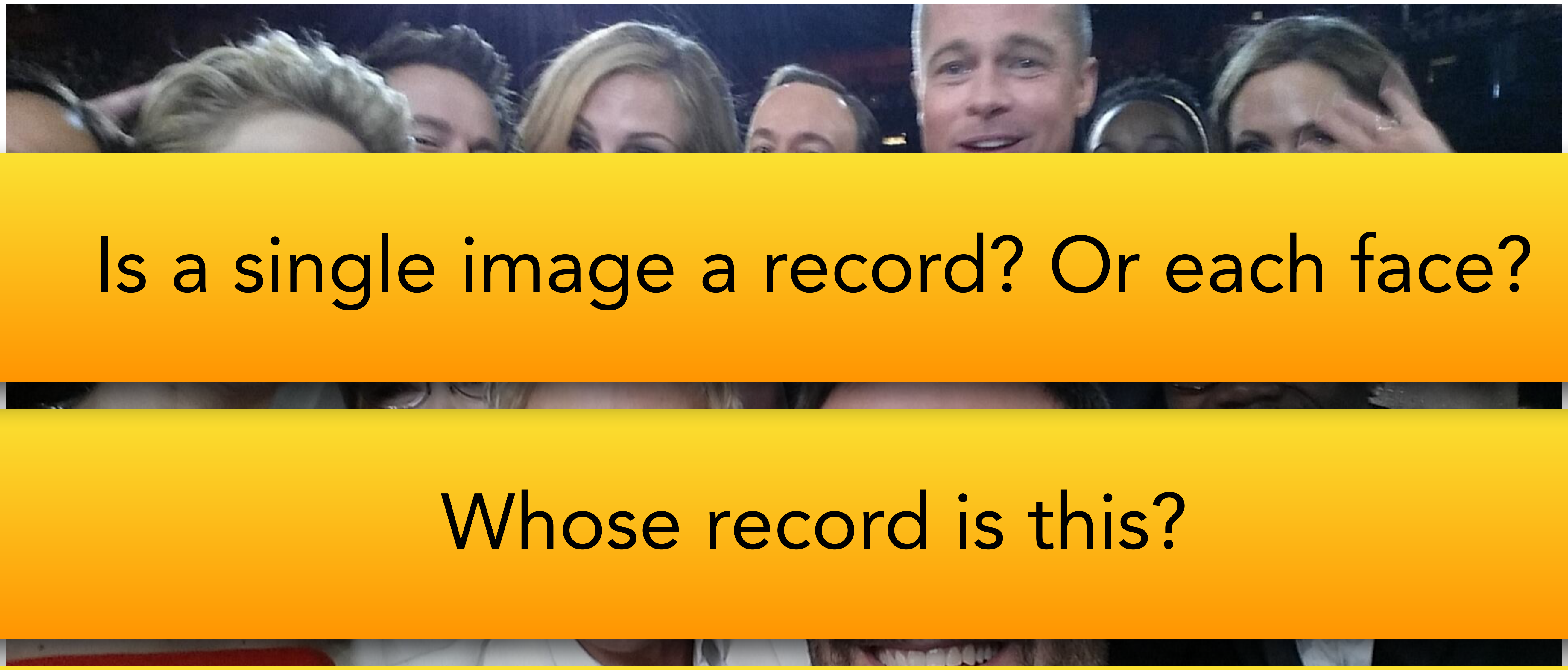


Is a single image a record? Or each face?

Whose record is this?

DP doesn't capture the nuances of privacy for text!

Or even other data-modalities! —images:



Conclusion

"So… Short Story long…"

Conclusion

- **What DP is:**
  - A great tool for computing **private statistics**, over independent **tabular data**
  - Context-free, **worst-case** privacy measure

Conclusion

- **What DP is:**
    - A great tool for computing **private statistics**, over independent **tabular data**
    - Context-free, **worst-case** privacy measure
- **What DP is not:**
    - **Free** in terms of data utility
    - A sensitive data/span **detection** and **scrubbing** tool

Thank You!

niloofar@cs.washington.edu

Also thanks to Katherine Lee, A. Feeder Cooper, Matthew Jagielski, Milad Nasr, Gautam Kamath and Yejin Choi for helpful feedback and discussions!