CSCI 699: Privacy-Preserving Machine Learning

Sai Praneeth Karimireddy

Agenda

01

Logistics

02

Why Privacy

03

What Privacy

04

Privacy in ML

Course Logisitics

- Class: Mon 4:00 to 7:30 pm, room: SGM 226
- Office hours: Wed 5:00 to 7pm, room: TBA
- Course website: spkreddy.org/ppmlfall2025.html
- Slack channel for QAs/discussion, assignment submission on
- Email: karimire@usc.edu (add CSCI 699 in subject)
- Anonymous feedback: https://forms.gle/8gta7KMHm2w3p1ro9

Course overview

- What even is privacy?
- How can you train a model while guaranteeing privacy?
- You say your training is safe, but how can I verify?
- I still don’t trust you with my data. Now what?
- What about copyright?
- How can the internet ad-economy function under GenAI?

Disclaimer

- The material we cover will be **hard.**

- **Diverse** topics and techniques, requires mathematical maturity.
  - **probability**
  - linear algebra
  - machine learning

- Cutting edge of ML research.

- Ideal outcome: you find a new question you are excited about and write a NeurIPS/ICML workshop-level paper. 3 last time - 1 by master students!

Grading

- 3 Assignments: **30%**
  - short: checking your understanding of the core concepts
  - practical: play with the concepts
- Project report: **35%** Due exam day, more details next.
- Paper reading and discussion: **35%**

Project Report: 35%

Option 1 (paper reading)

- Team up with others who signed up for similar papers - 1 to 3.
- Teach each other your papers and related background.
- Replicate the core experiments of SOTA
- Write up a 4 page report.

Option 2 (research - encouraged)

- Teams of 1-3.
- Come up with an research question (based on what you've read or otherwise)
- Setup a meeting to get my feedback **before Oct 6 (fall break).**
- Write up a 4 page report.

Paper Reading & Discussion: 35%

We will use a **Role-Playing** disucussion format.

- Each week post fall break, we will discuss 2-3 papers.
- Everyone picks one of the following roles for each:
  - **Presenter*:** present the paper
  - **Antagonist:** find flaws, missing experiments
  - **Archaeologist:** effect of this paper on the field
  - **Researcher**: abstract of a pretend follow-up paper
  - **Practioner:** turn into a product and pitch it
- 1 presenter per paper - in class presentation**: 20%**
- Rest, split among 4 roles. Submit 1 paragraph before class on brightspace. Discuss in class. **15%**
- Everyone takes all roles equally

Schedule

Week	Date	Topic (lecture)	Presentation	Due
1	Aug 25	Course logistics, Why privacy, attempts at privacy, linkage atta		
	~~Sep 1~~	~~Labor day~~		
2	Sep 8	Hypothesis testing, Laplace mechanism, properties of DP, Gau		**HW 1 due**
3	Sep 15	Approximate DP, advanced composition, GD, DP-GD, SGD, DP-S		
4	Sep 22	f-DP, Gaussian DP, privacy auditing		**HW 2 due**
5	Sep 29	membership inference attacks, Privacy auditing		
6	Oct 6	Copyright, memorization, watermarking		**HW 3 due, Project topic**
7	Oct 13	Data attribution	reconstruction attacks, LIRA membership inference attacks	papers 1-3
8	Oct 20	Unlearning	measuring memorization	papers 4-6
9	Oct 27		data attribution and watermarking	papers 7-9
10	Nov 3		unlearning	papers 10-12
11	Nov 10	Local DP, decentralized privacy, federated learning	privacy in LLMs	papers 13-15
12	Nov 17		Sanitization approaches, prompt defenses contextual integrit	papers 16-18
13	Nov 24		Local DP, decentralized privacy	papers 19-21
14	Dec 1		federated privacy & law	papers 21-24
	~~Dec 8~~	~~Study break~~		
	Dec 15			**Project report due**

Agenda

01

Logistics

02

Why Privacy

03

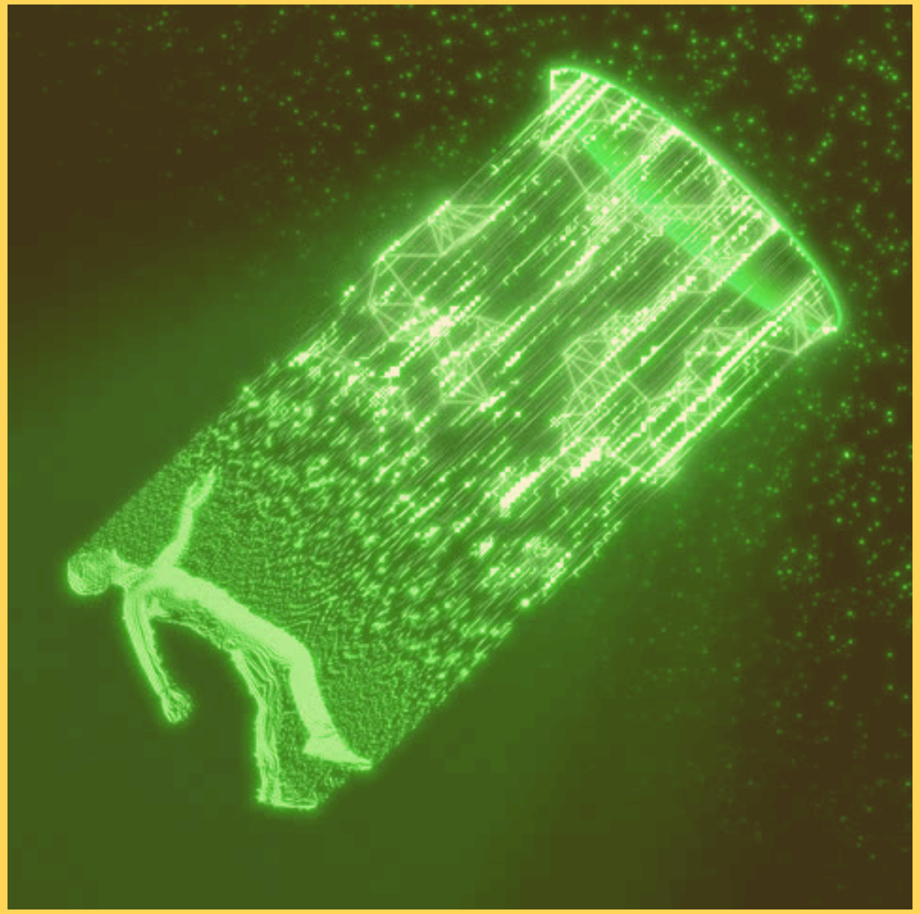
What Privacy

04

Privacy in ML

"The world's most valuable resource is no longer oil, but data. " - Economist, 2017

Tesla, Uber, Dominos are data companies.

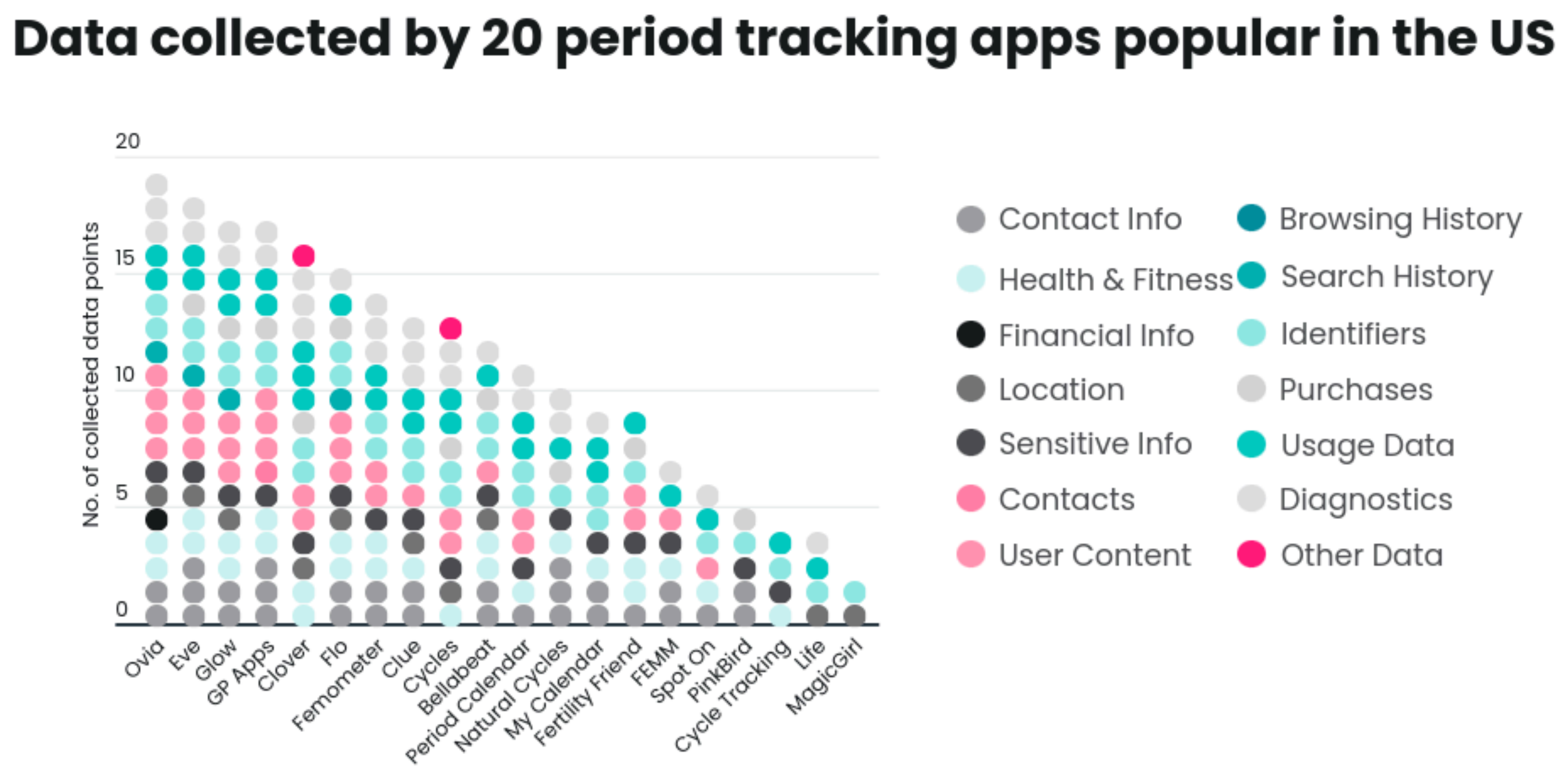


src: @perfectloop
used with permission

Why privacy?


1999
2010
“No one cares”
“If you’ve done nothing wrong…”
2009
Google

Why privacy? Case 1



Surfshark 2022

- Menstrual tracking apps track a ton of data.
- They, like many other apps, sell data to **data brokers.**
- Can infer pregnancy and abortions. Illegal in a large part of US.
- "Wrong" according to who?

Why privacy? Case 2

NY Times 2019

- Apps also sell your location to data brokers
- Anyone can buy it. Lots of people do.
- Easily identify protestors and trace people to homes
- Senior Defense Department official and his wife identified at the Women's March.

Why privacy? Case 3

U.S. NEWS

View site information

23andMe user data targeting Ashkenazi Jews leaked online

A database that has been shared on dark web forums and viewed by NBC News has a list of 999,999 people who allegedly have used the service.

Will you share my data with my insurance company or my employer?

No. Your data (genetic or self-reported) will not be provided to an insurance company or employer. End of story.

- You don't know who can use that data for what purpose.
- Datasets can get hacked and leaked.
- Companies (and their data) can get bought and sold.
- 23&Me paid $30M for a data breach, went bankrupt. What happens to user data?

Why privacy? Summary

The New York Times

As Troops Walk the Streets, Washington Restaurants Report a Slump

When the chef Rock Harper read the news this month that President Trump was going to send the National Guard into the streets of Washington,...

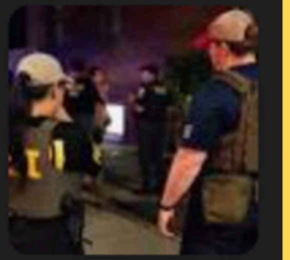
2 days ago

The Washington Post

'The city is dead': D.C. restaurant reservations drop amid federal crackdown

August restaurant week can be a lifeline for D.C. food businesses, but high-visibility arrests and heavy law enforcement presence are...

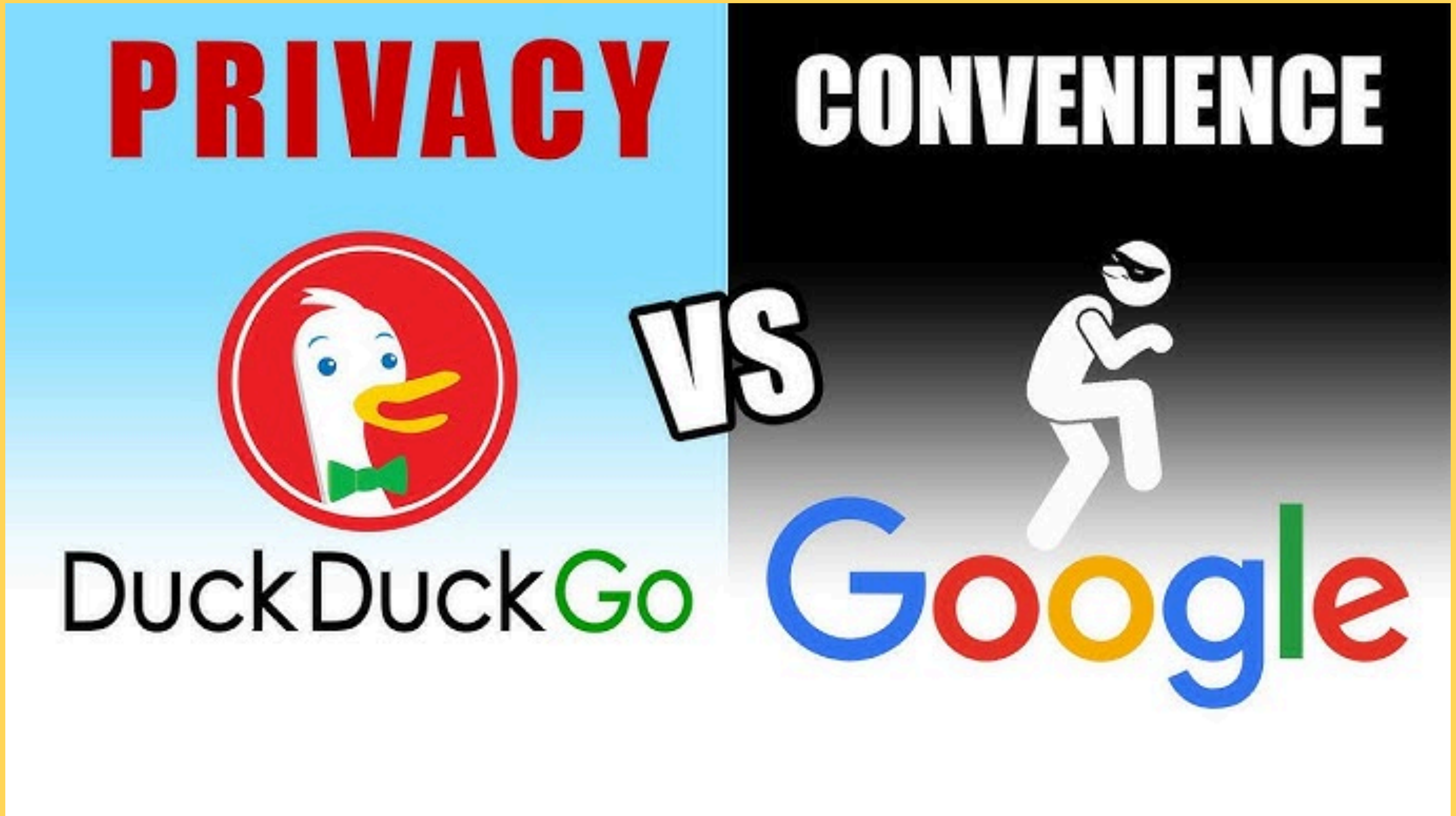
5 days ago



EU Law analysis 2020

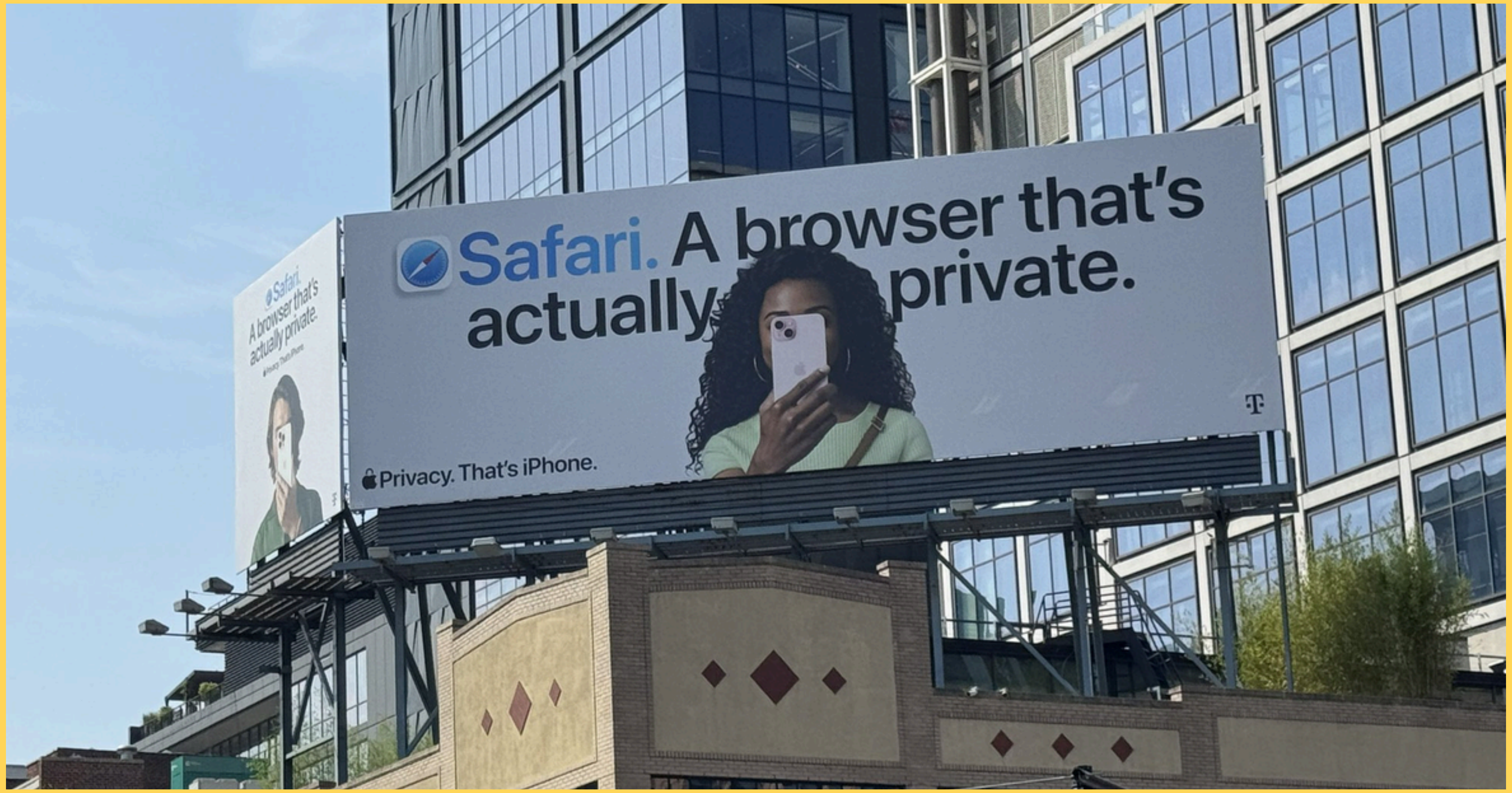
- You are being looked at, but you can't look back.
- If a flag is raised, very expensive to deal with.
- You will change your behavior to be overly cautious and not raise flags => "chilling effect"
- Privacy is about power-imbalance.

Privacy is also BIG BUSINESS

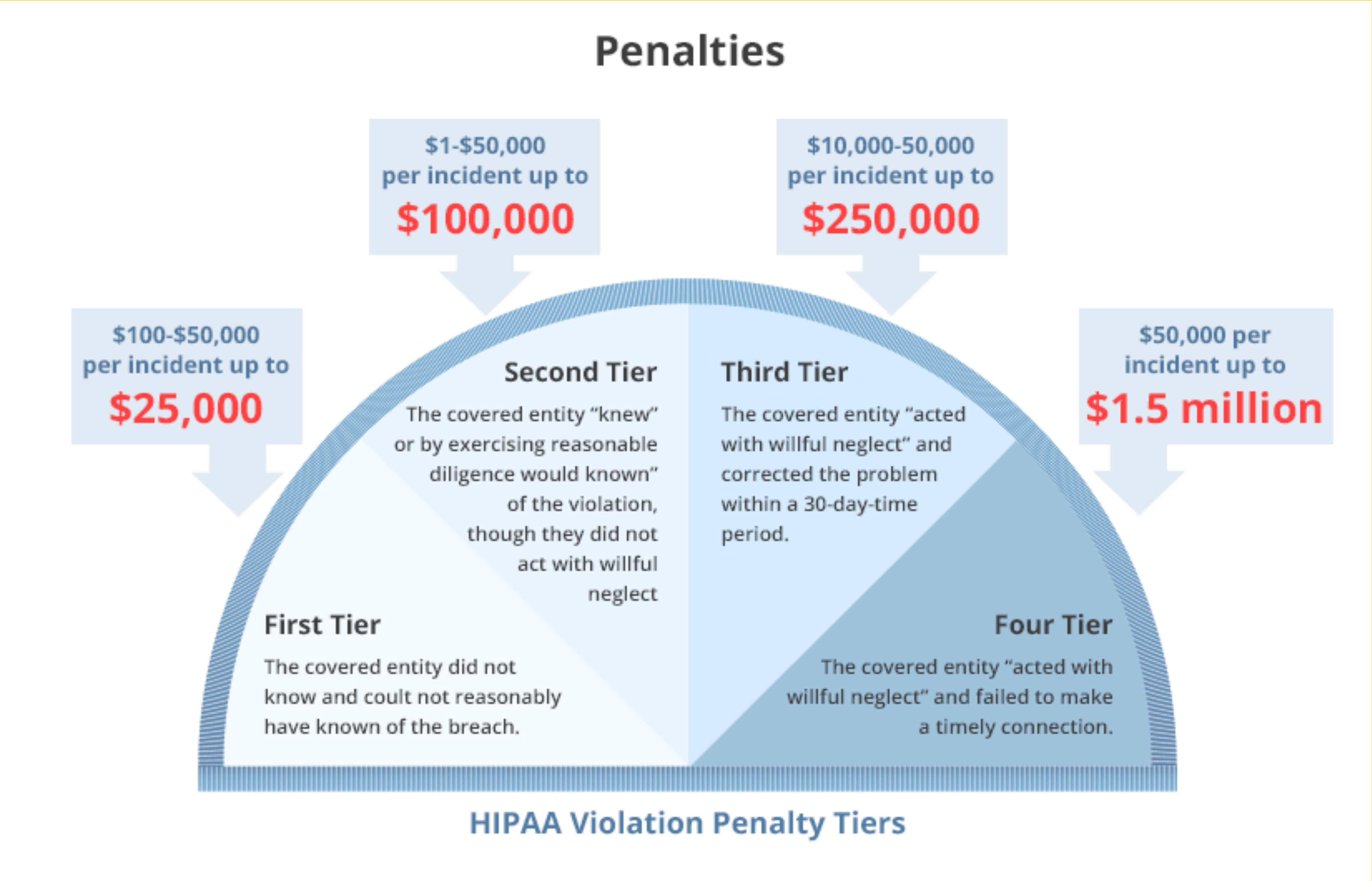


- If you don't trust Google, you may start using alternatives
- Google will lose out!
- Lots of effort in ensuring baseline trust and privacy.

Privacy is also BIG BUSINESS



Privacy is also BIG BUSINESS



- HIPAA violation fines of $5 million in 2023
- 2022 GDPR fines were $2 billion!

But what is “privacy”?

“Data People” by Jamillah Knowles

But what is “privacy”?

“Data People” by Jamillah Knowles

- “The right to be let alone” - Warren II & Justice Louis Brandeis.
- To exercise your other rights freely without coercion, influence, or persuasion.
- No really. what is privacy?

Agenda

01

Logistics

02

Why Privacy

03

What Privacy

04

Privacy in ML

De-identification

1 Name	10 Licence details
2 Phone Number	11 VIN (Vehicle Identification Number)
3 Dates (admission date, discharge date, appointment date etc.)	12 Identifiers in Medical devices (Pacemaker)
4 Fax details	13 Website URLs
5 Email ID	14 IP Address
6 SSN (Social Security Number)	15 Biometrics (Fingerprint)
7 MRN (Medical Record Number)	16 Full-face photographs or images with differentiators (facial scars, moles etc.)
8 HPBN (Health Plan Beneficiary Number)	17 Any other unique identifiers
9 Medical Certificates	18 Address (if it has information on the city, street, and house number)

- Remove “sensitive” and “private” attributes: Personally Identifiable Information (PII)
- HIPAA identifies 18 attributes which if present would make the data PHI: Private Health Information.
- Note number 17

De-identification

PATIENT
[redacted]

DOB	[redacted]
AGE	46 yrs
SEX	Female
PRN	[redacted]

FACILITY
[redacted] M.D.
T [redacted]
F [redacted]
[redacted]

Patient identifying details and demographics

FIRST NAME	[redacted]	SEX	Female	ETHNICITY	Hispanic or Latino
MIDDLE NAME	-	DATE OF BIRTH	[redacted]		
LAST NAME	[redacted]	DATE OF DEATH	-	PREF. LANGUAGE	English
SSN	[redacted]	PRN	[redacted]	RACE	White
				STATUS	Active patient

CONTACT INFORMATION

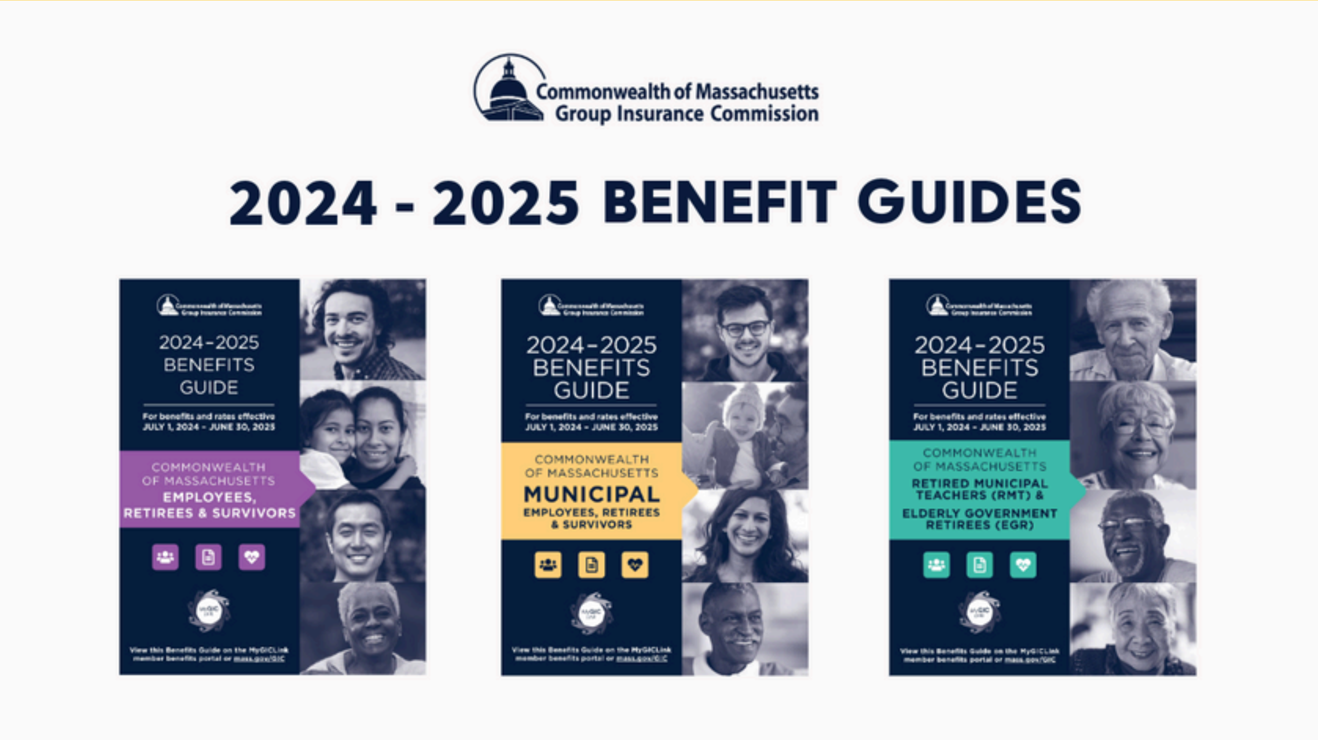
ADDRESS LINE 1	[redacted]	CONTACT BY	Mobile Phone
		EMAIL	[redacted]
ADDRESS LINE 2	-		
CITY	[redacted]	HOME PHONE	-
STATE	[redacted]	MOBILE PHONE	[redacted]
ZIP CODE	[redacted]	OFFICE PHONE	-
		OFFICE EXTENSION	-

FAMILY INFORMATION

NEXT OF KIN	[redacted]	PATIENT'S MOTHER'S MAIDEN NAME	-
RELATION TO PATIENT	Friend		
PHONE	[redacted]		
ADDRESS	-		

- A lot of work!
- But are we good?

De-identification

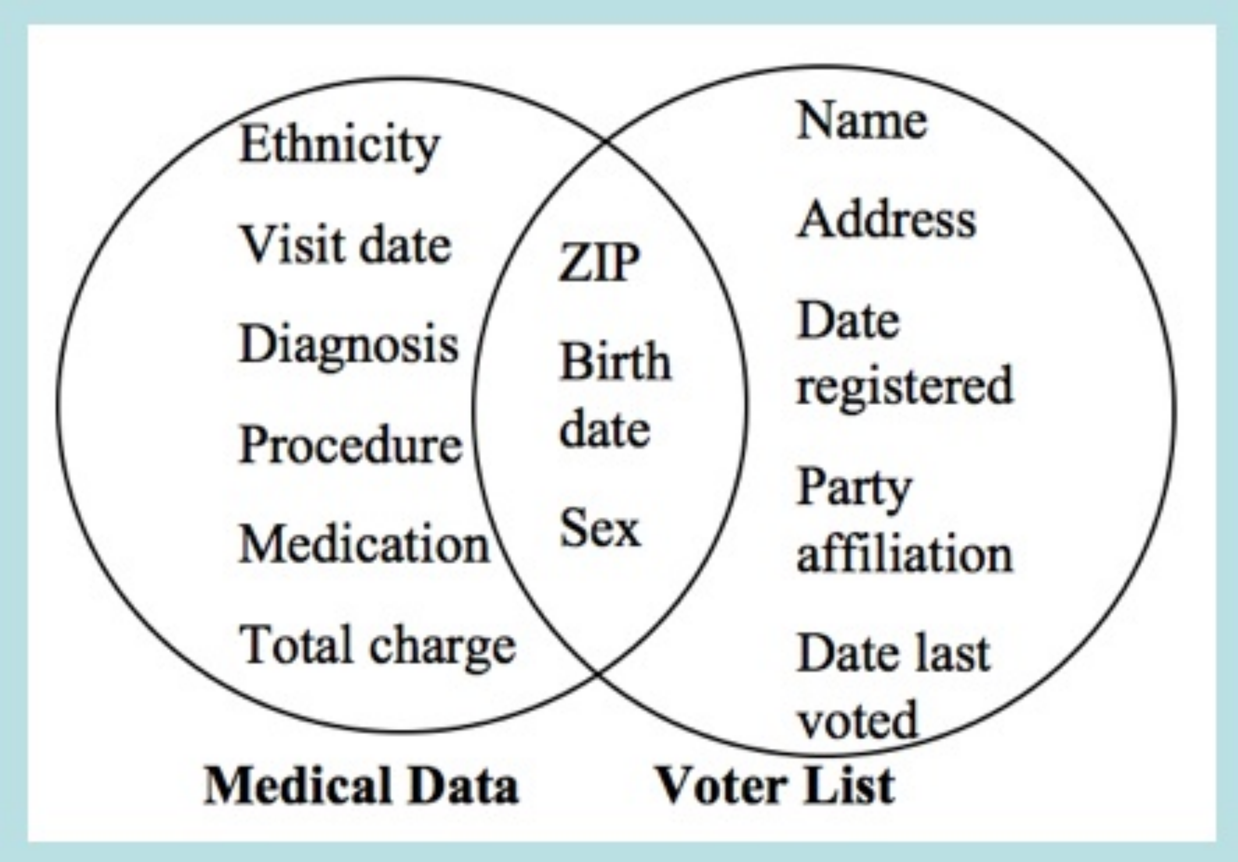



Bill Weld

- GIC released "anonymized" data on state employees that showed every single hospital visit to researchers.
- Bill Weld assured the public that GIC had protected patient privacy by deleting identifiers.
- They still had DOB, ZipCode, Sex, along with hospital visits, diagnosis.

De-identification

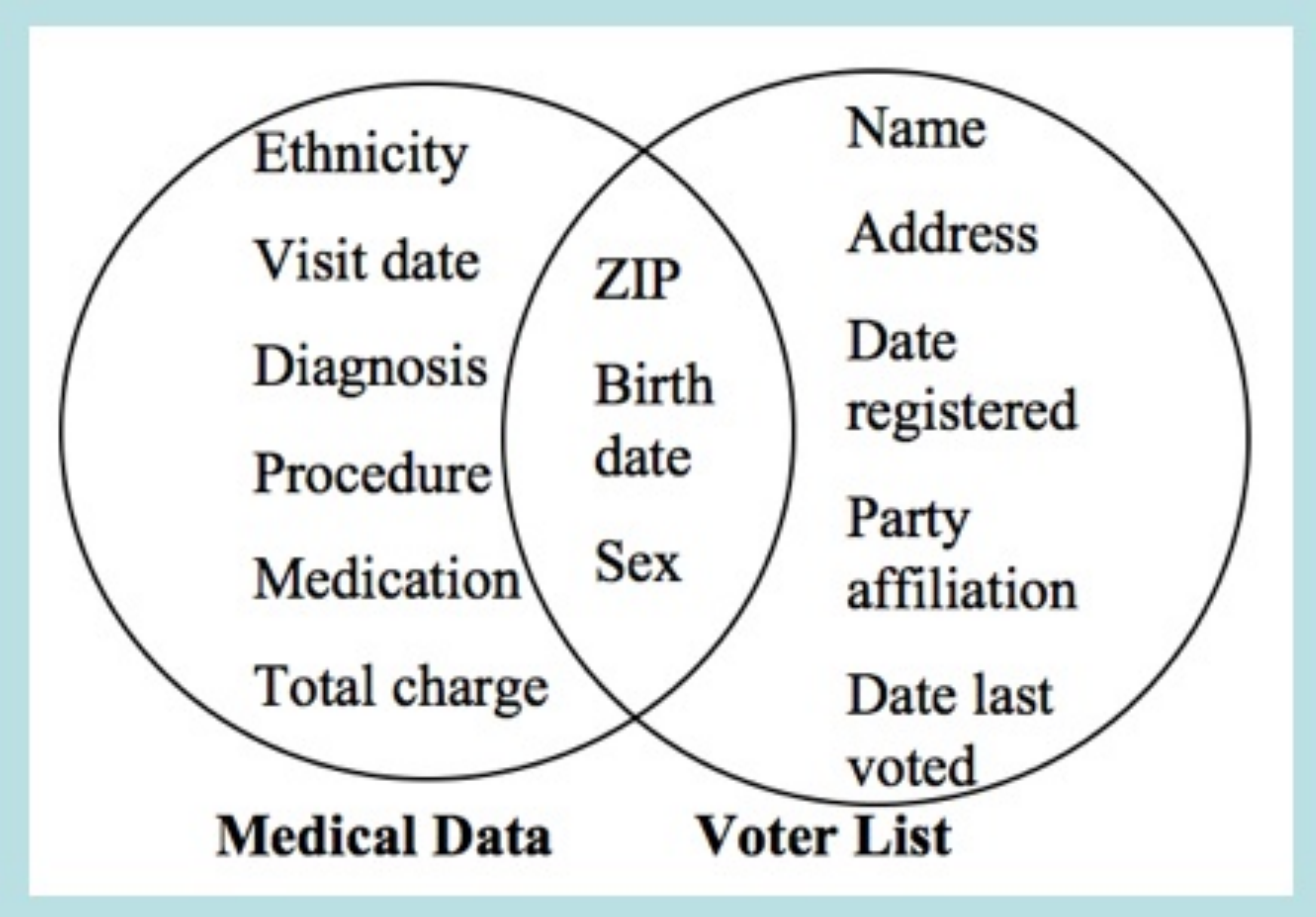
Latanya Sweeney 1997: 87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}



Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code.

K-anonymity

Sweeney 1997: 87% of the U.S. Population are **uniquely identified by {date of birth, gender, ZIP}**



What if there were 10 others who had the exact same attributes as Bill?

K-anonymity

Definition [Sweeny 1998]: For every row in the database, there should be (k-1) others with the exact same attributes.

Name	DoB	Gender	Height (cm)	Weight (kg)	Address	Disease
Jenna Wilson	1949-04-23	Male	166	117	6639 Mayo Crescent Suite 839, South Austin, VT 27102	Heart Disease
Anita Garcia	1950-02-02	Male	152	75	9674 Ann Ways, Fullerborough, UT 74286	Asthma
Sheila Ramirez	1980-08-04	Female	175	114	39357 White Island Suite 518, Kathystad, LA 31540	Diabetes
Ryan Jensen	1998-03-10	Male	174	94	31039 Duncan Glens Suite 244, South Annahaven, CA 38497	Heart Disease
Edward Lewis	1974-11-01	Male	157	88	USNS Butler, FPO AP 27077	Asthma
Jared Knight	1957-08-13	Female	183	99	860 Nichols Summit Suite 235, North Tina, CA 24369	Obesity

K-anonymity: supression

Name	DoB	Gender	Height (cm)	Weight (kg)	Address	Disease
Jenna Wilson	1949-04-23	Male	166	117	6639 Mayo Crescent Suite 839, South Austin, VT 27102	Heart Disease
Anita Garcia	1950-02-02	Male	152	75	9674 Ann Ways, Fullerborough, UT 74286	Asthma
Sheila Ramirez	1980-08-04	Female	175	114	39357 White Island Suite 518, Kathystad, LA 31540	Diabetes
Ryan Jensen	1998-03-10	Male	174	94	31039 Duncan Glens Suite 244, South Annahaven, CA 38497	Heart Disease
Edward Lewis	1974-11-01	Male	157	88	USNS Butler, FPO AP 27077	Asthma
Jared Knight	1957-08-13	Female	183	99	860 Nichols Summit Suite 235, North Tina, CA 24369	Obesity

K-anonymity: generalization

~~DoB~~	Gender	Height (cm)	Weight (kg)	Disease
~~1949-04-23~~	Male	166	117	Heart Disease
~~1950-02-02~~	Male	152	75	Asthma
~~1980-08-04~~	Female	175	114	Diabetes
~~1998-03-10~~	Male	174	94	Heart Disease
~~1974-11-01~~	Male	157	88	Asthma
~~1957-08-13~~	Female	183	99	Obesity

K-anonymity: generalization

Age	Gender	Height (cm)	Weight	Disease
45-65	Male	160-180	Normal	Heart Disease
45-65	Male	140-160	Normal	Asthma
25-45	Female	160-180	Normal	Diabetes
45-65	Male	160-180	Normal	Heart Disease
45-65	Male	140-160	Normal	Asthma
65+	Female	180-200	Overweight	Obesity

K-anonymity: outlier removal

Age	Gender	Height (cm)	Weight	Disease
45-65	Male	160-180	Normal	Heart Disease
45-65	Male	140-160	Normal	Asthma
~~25-45~~	~~Female~~	~~160-180~~	~~Normal~~	~~Diabetes~~
45-65	Male	160-180	Normal	Heart Disease
45-65	Male	140-160	Normal	Asthma
~~65+~~	~~Female~~	~~180-200~~	~~Overweight~~	~~Obesity~~

Satisfies 2-anonymity

K-anonymity

- are Strava heatmaps de-identified?
- Do they satisfy k-anonymity?
- What went wrong?

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- Latest: Strava suggests military users 'opt out' of heatmap as row deepens

A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

ℓ-diversity

ℓ-Diversity: Privacy Beyond k-Anonymity

Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer
Muthuramakrishnan Venkitasubramaniam
Department of Computer Science, Cornell University
{mvnak, johannes, dkifer, vmuthu}@cs.cornell.edu

Definition: For each set of attributes, make sure there are at diverse (least l) sensitive attributes.

ℓ-diversity

Age	Gender	Height (cm)	Weight	Disease
45-65	Male	160-180	Normal	Heart Disease
45-65	Male	140-160	Normal	Asthma
45-65	Male	160-180	Normal	Heart Disease
45-65	Male	140-160	Normal	Asthma

Definition: For each set of attributes, make sure there are at diverse (least l) sensitive attributes.
Is our 2-anonymous table 2-diverse? Can we make it?

Lots of back and forth

t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity

Ninghui Li, Tiancheng Li
Department of Computer Science, Purdue University
{ninghui, li83}@cs.purdue.edu

Suresh Venkatasubramanian
AT&T Labs – Research
suresh@research.att.com

Hiding the Presence of Individuals from Shared Databases

M. Ercan Nergiz *
CS Dept., Purdue University
305 N. University Street
West Lafayette, Indiana,
47907-2107
mnergiz@cs.purdue.edu

Maurizio Atzori †
KDD Laboratory, ISTI-CNR
Area della ricerca di Pisa
via G. Moruzzi 1
56124 Pisa, Italy
atzori@di.unipi.it

Christopher W. Clifton
CS Dept., Purdue University
305 N. University Street
West Lafayette, Indiana,
47907-2107
clifton@cs.purdue.edu

...

Lots of back and forth. even recently. privacy is HARD.

[Submitted on 6 Oct 2020 (v1), last revised 24 Feb 2021 (this version, v2)]

InstaHide: Instance-hiding Schemes for Private Distributed Learning

Yangsibo Huang, Zhao Song, Kai Li, Sanjeev Arora

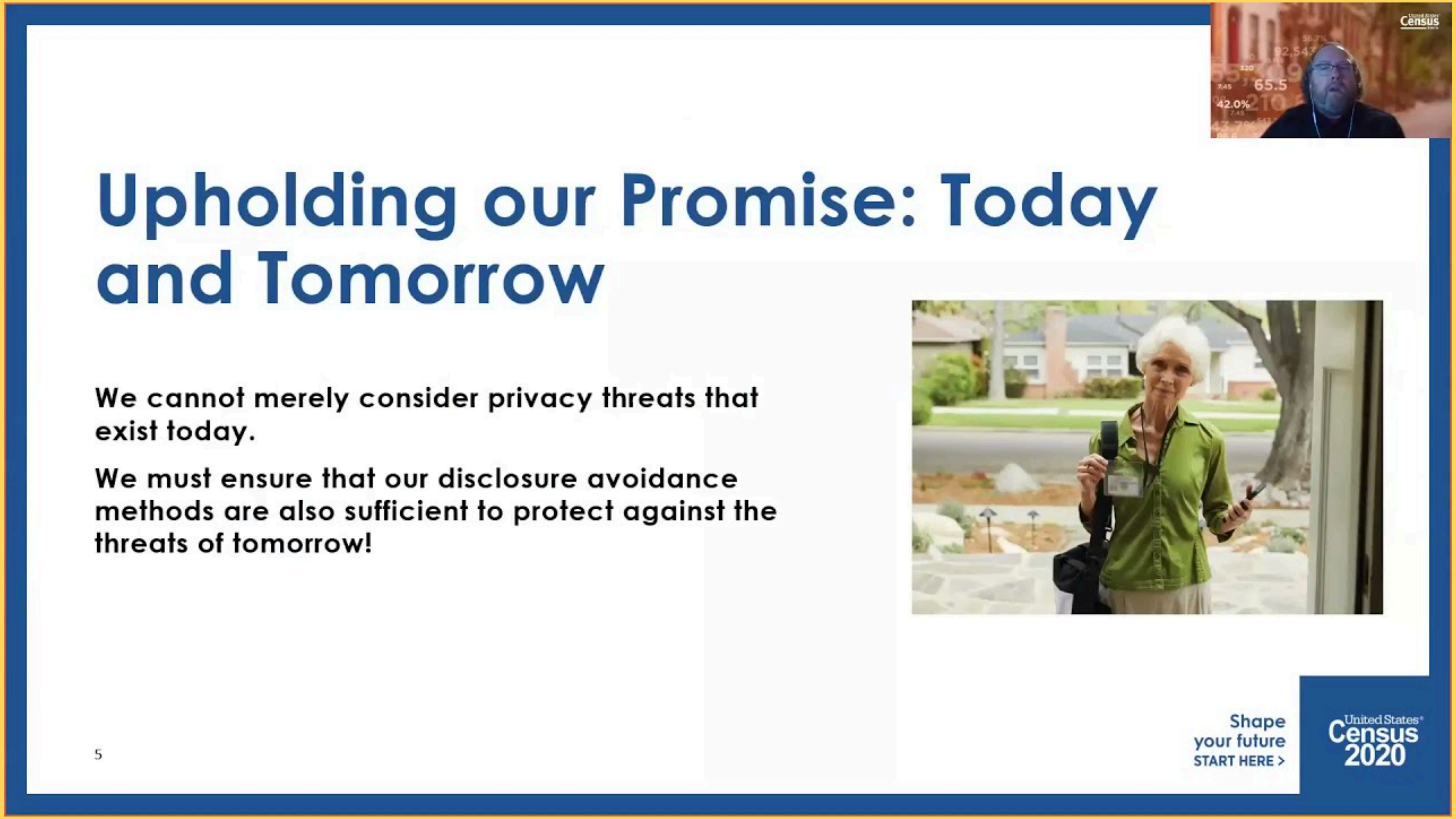
InstaHide Disappointingly Wins Bell Labs Prize, 2nd Place

by **Nicholas Carlini** 2020-12-05

Is Private Learning Possible with Instance Encoding?

Nicholas Carlini ncarlini@google.com	Samuel Deng sd3013@columbia.edu	Sanjam Garg sanjamg@berkeley.edu
Somesh Jha jha@cs.wisc.edu	Saeed Mahloujifar sfar@princeton.edu	Mohammad Mahmoody mohammad@virginia.edu
Shuang Song shuangsong@google.com	Abhradeep Thakurta athakurta@google.com	Florian Tramèr tramer@cs.stanford.edu

Where to go? Differential Privacy



Quantifying Privacy Leakage

Recap

- We saw many definitions of privacy
    - De-identification / suppression
    - K-anonymity
    - L-diversity
- We saw none of them really protected privacy and were easily broken
- Hinted at a more widely accepted definition.

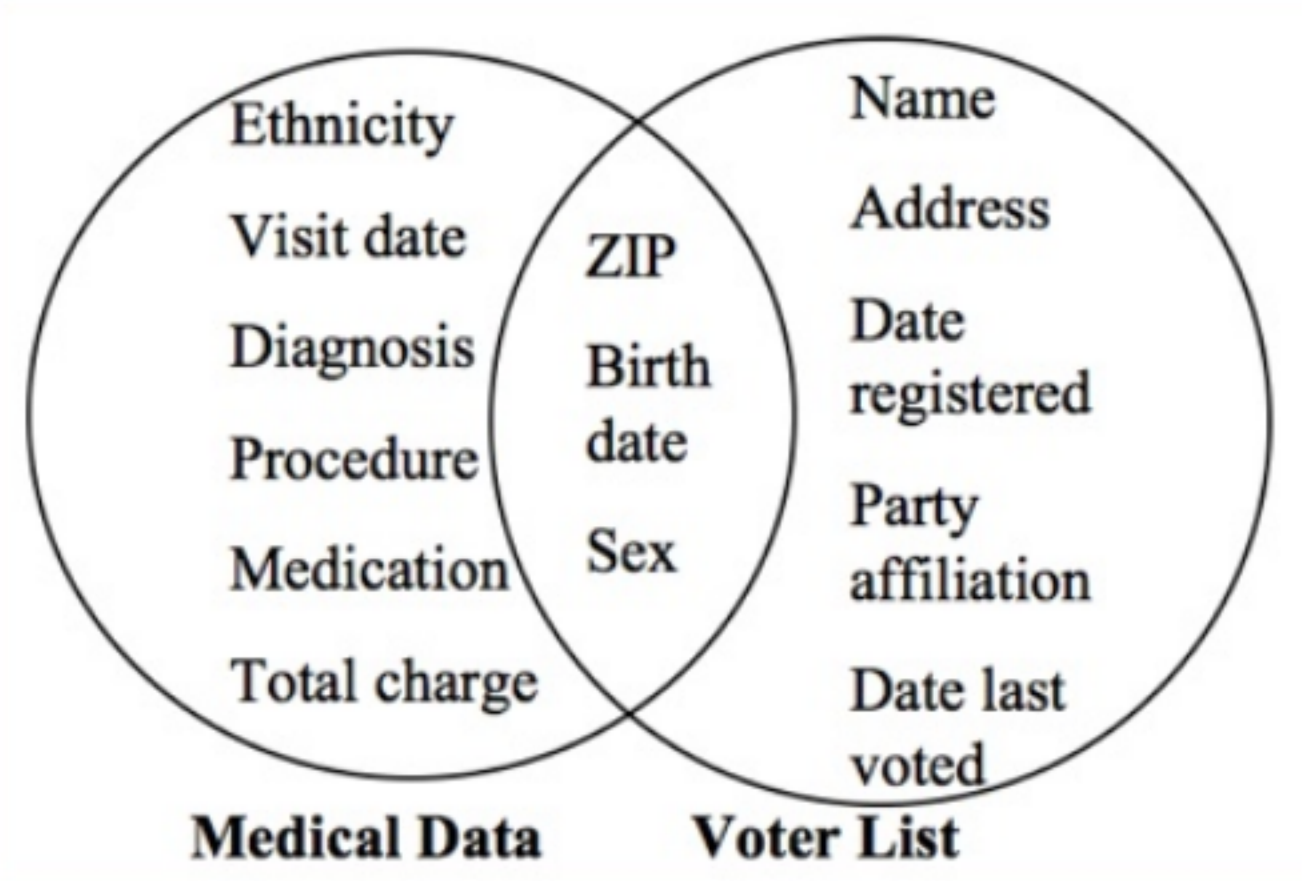
Takeaways

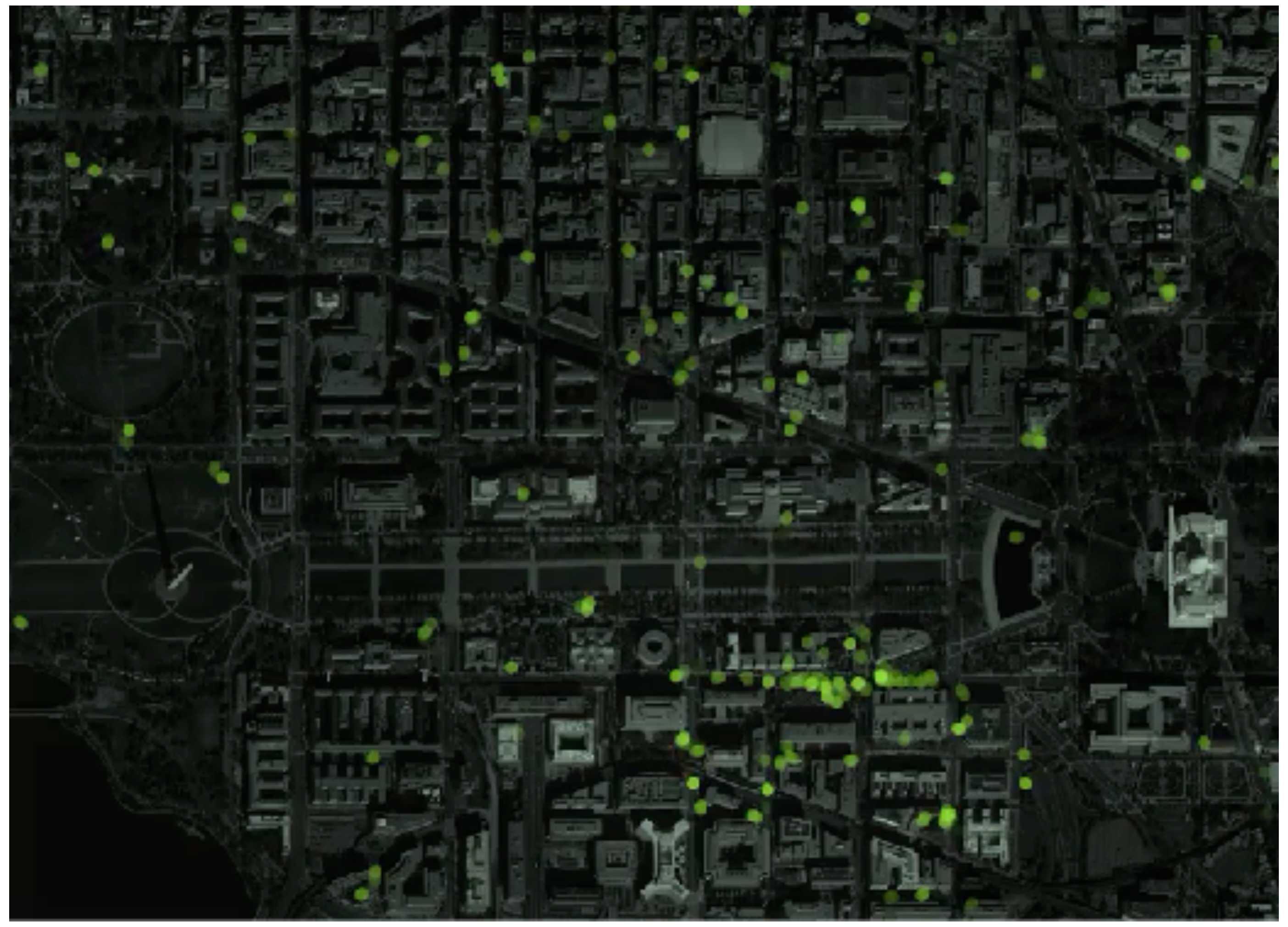
Requirements for privacy definition

- Unaffected by auxiliary information: we should not be able to combine extra data to undo privacy.

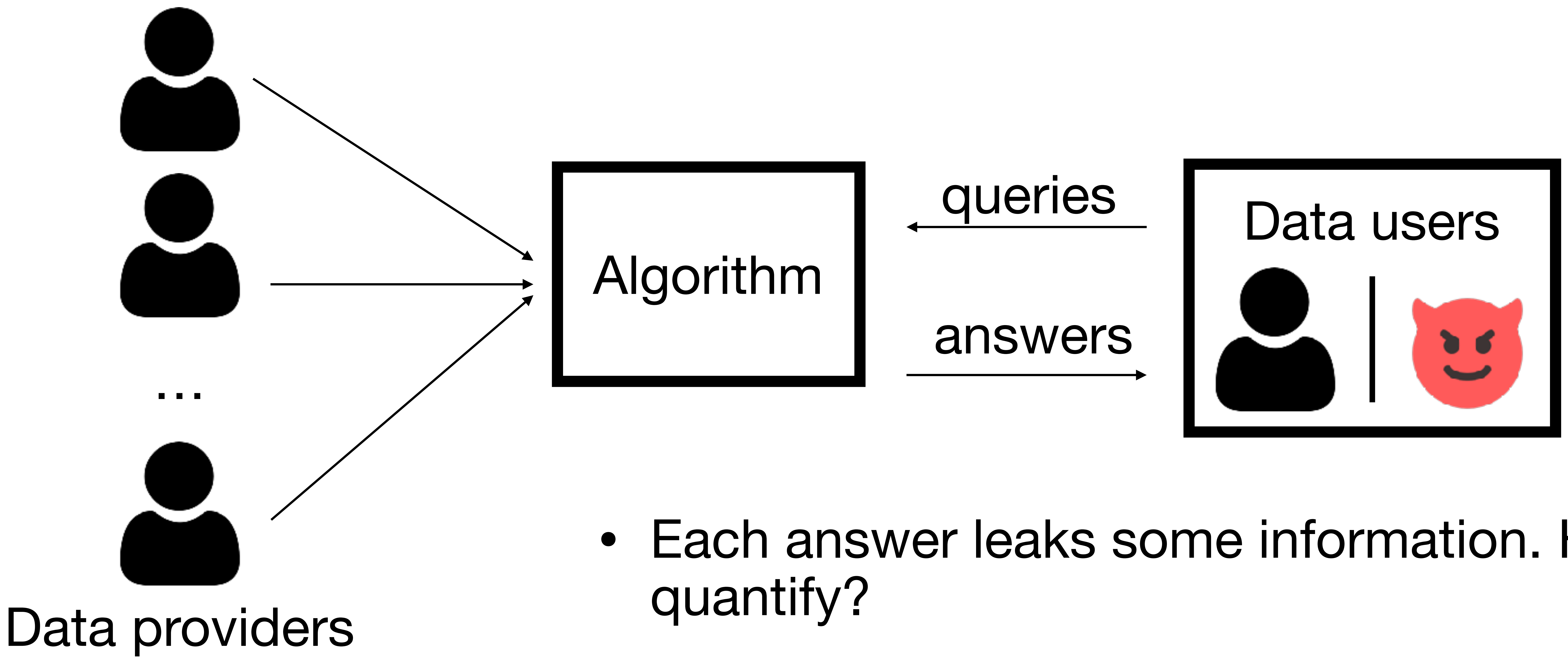
- Composition: We should understand what happens when data is continuously released.

- Today we will come with such a privacy definition.





Goals of PPML



- Each answer leaks some information. How to quantify?
- How to balance usefulness of answers vs. privacy being leaked?

Quantifying Privacy Leakage

Attempt 1

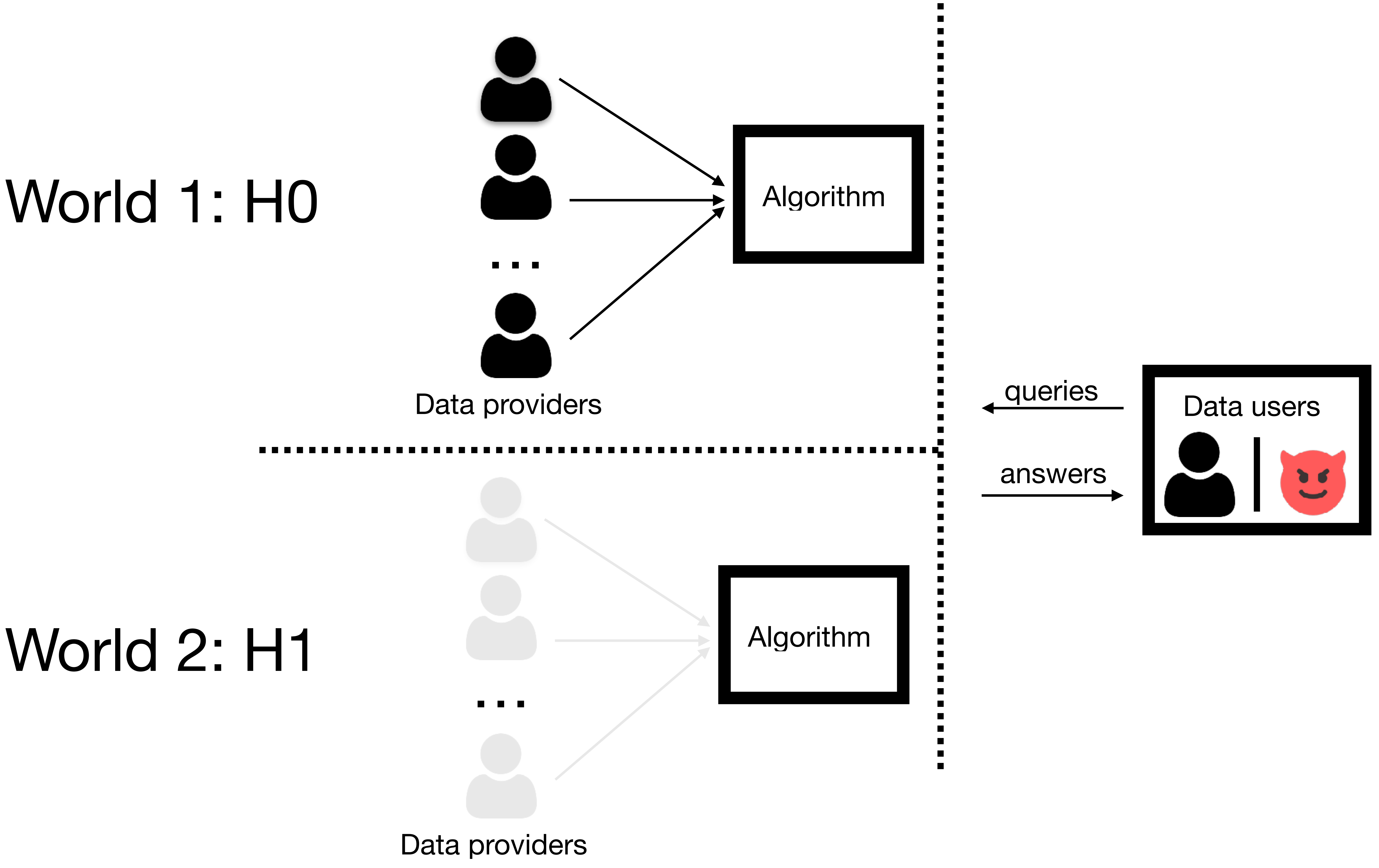
Absolute Privacy: quantify **total** information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- Bayesian version: the posterior and prior are identical

Quantifying Privacy Leakage

Attempt 1



We are either in world 1 or world 2. The adv cannot tell which world we are in.

Quantifying Privacy Leakage

Attempt 1

Absolute Privacy: quantify **total** information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- **Problem 1**: Impossible to reveal anything useful about data since any useful answer will provide some previously unknown information.

Quantifying Privacy Leakage

Attempt 1: Problems

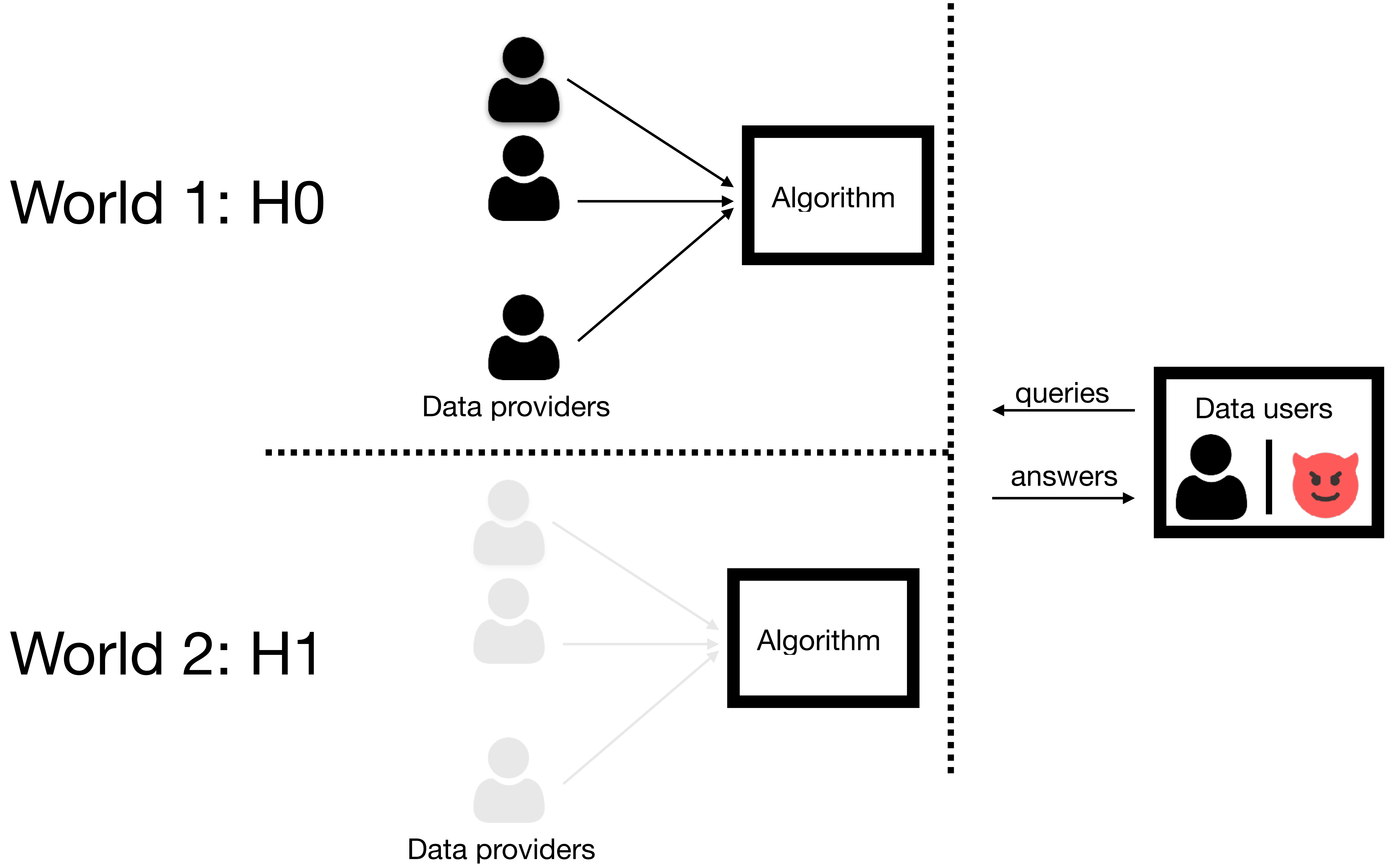
Absolute Privacy: quantify **total** information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- **Problem 2**: What I know before changes with auxiliary information.
- Did the model leak information about Bob?
  - Bob is a smoker, but his data was not used to train the model.
  - The model said smokers have higher risk of disease.
  - Bob's insurance premiums were raised.

Quantifying Privacy Leakage

Attempt 1: Problems



Any information about the distribution reveals which world we are in.

Quantifying Privacy Leakage

Attempt 1: Problems

Absolute Privacy: quantify **total** information leaked

"An answer to a query is private if the response reveals no more than was already known about the individuals in the data"

- **Problem 2**: What I know before changes with auxiliary information.
- We want to safeguard individual information (privacy) while revealing distributional/aggregate information (utility)

Quantifying Privacy Leakage

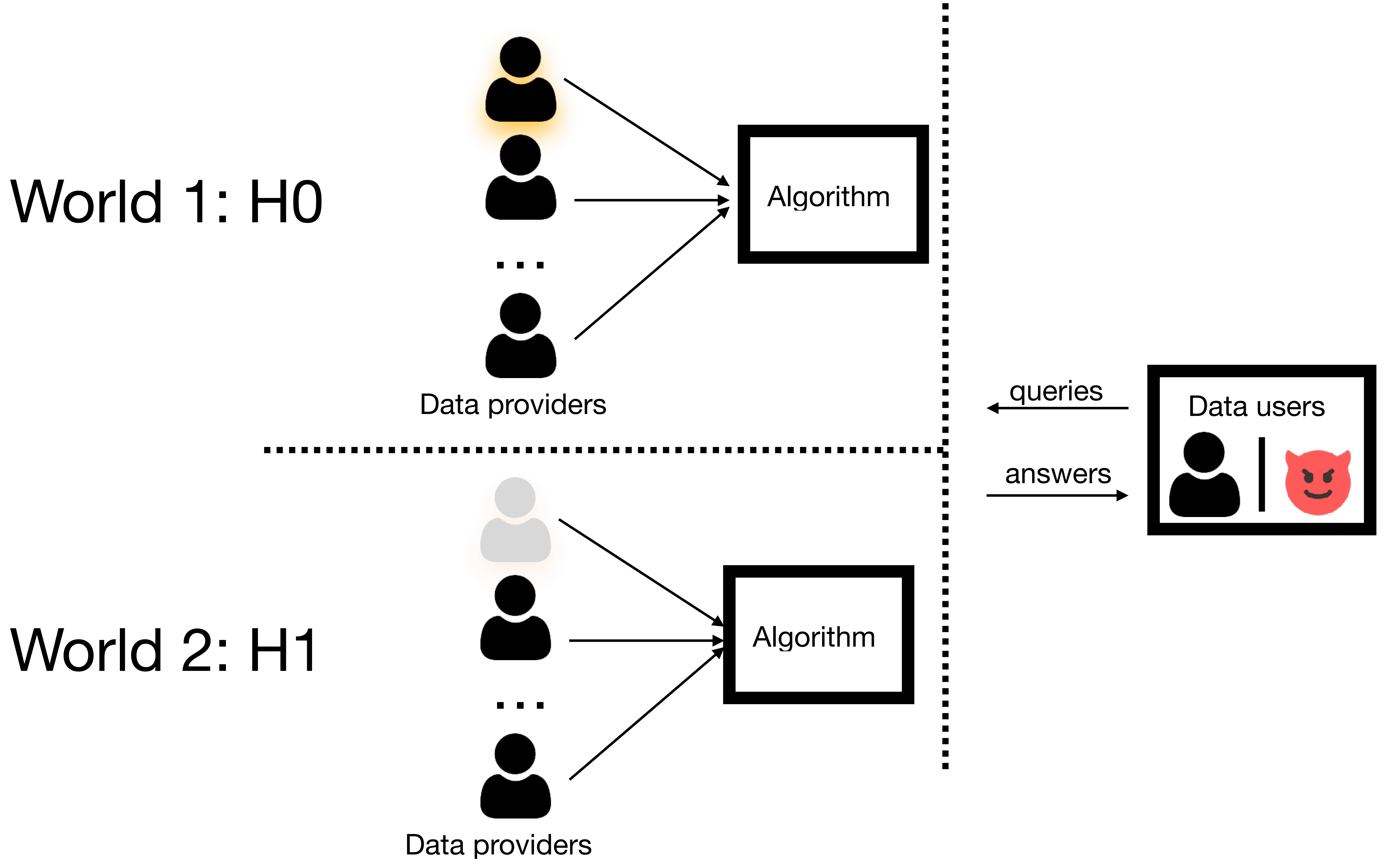
Attempt 2

Relative Privacy: quantify **individual** information leaked

"An analysis of a dataset is private if what can be learned about an individual in the dataset is not much more than what would be learned if the same analysis was conducted without them in the dataset"

Quantifying Privacy Leakage

Attempt 2



- In world 2 only Bob is removed/replaced.
- Now from the answer, how easily can guess the correct world?

Quantifying Privacy Leakage

Attempt 2

Relative Privacy: quantify **new** information leaked

"An analysis of a dataset is private if what can be learned about an individual in the dataset is not much more than what would be learned if the same analysis was conducted without them in the dataset"

- **Intuition**: Whether Bob is present in the data or not, the answer should not change much.
- Then, from looking at the answer, we will not learn whether Bob was present in the data or not.
- Gives Bob plausible deniability.

Aside: how is Putin's popularity calculated?

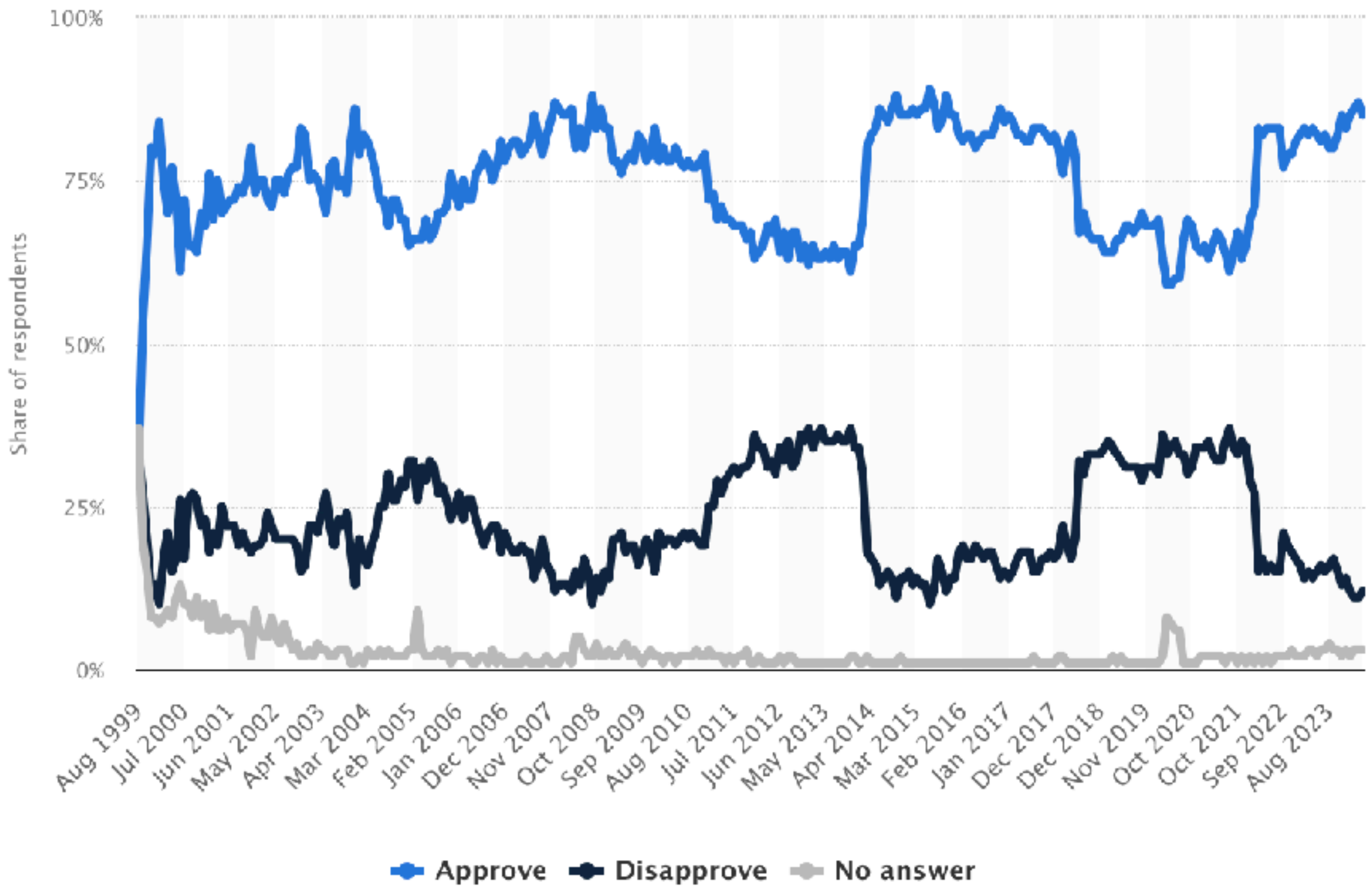
Plausible deniability as privacy

Poll: Russians Still Like Putin and Back the Ukraine War – but Are Anxious at Home

Most Russian survey respondents see the war in Ukraine as a broader conflict with the West and support it amid concerns about their own country's economy.

By Elliott Davis Jr. | Jan. 9, 2024

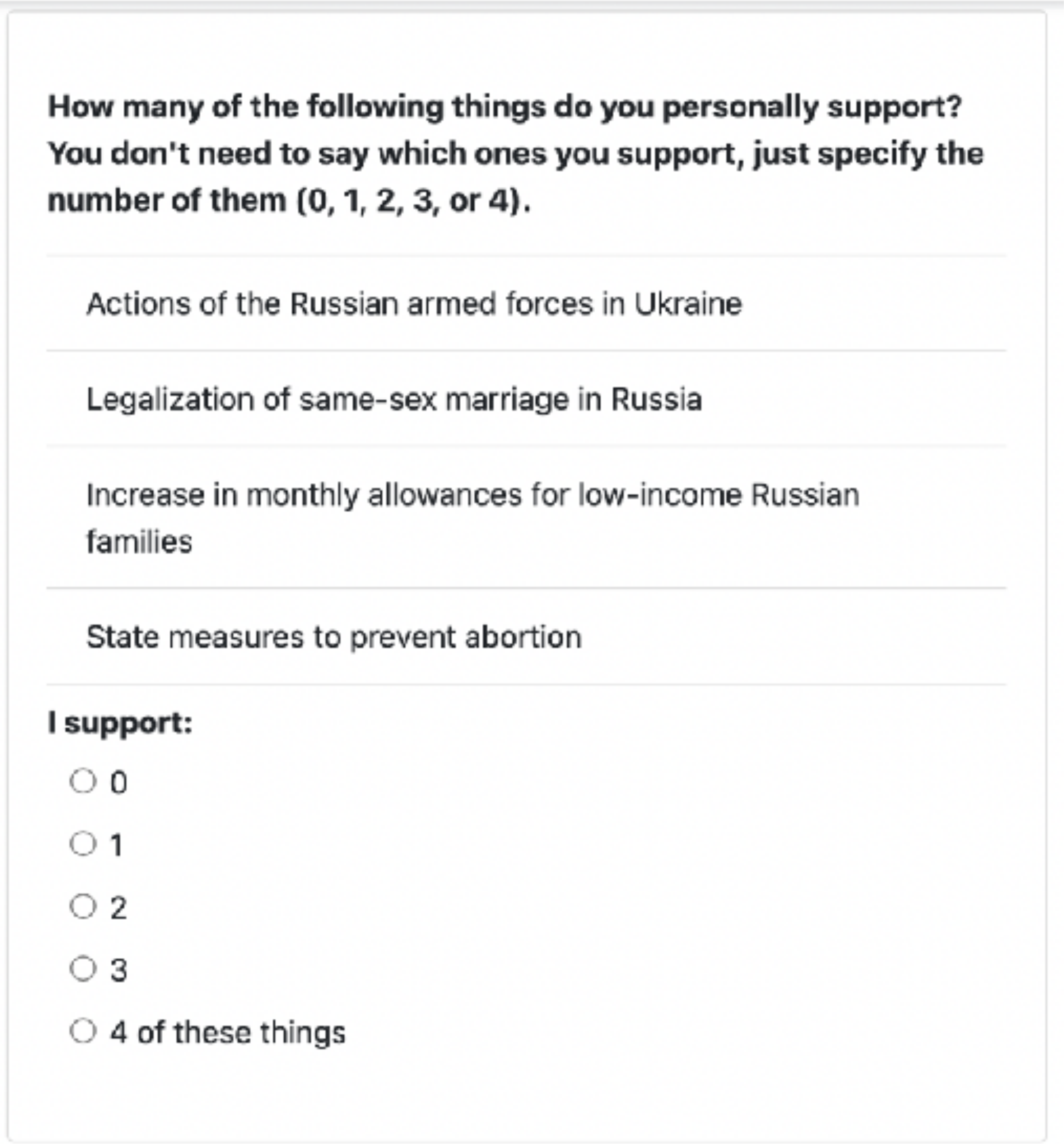
Do you approve of the activities of Vladimir Putin as the president (prime minister) of Russia?



Aside: how is Putin's popularity calculated?

List Experiment

- Split users randomly into two groups
- Design a set of options very similar to the one you actually care about
- To control only ask about the rest. To the treatment include your option.
- Does this confer plausible deniability?



How many of the following things do you personally support? You don't need to say which ones you support, just specify the number of them (0, 1, 2, 3, or 4).

Actions of the Russian armed forces in Ukraine

Legalization of same-sex marriage in Russia

Increase in monthly allowances for low-income Russian families

State measures to prevent abortion

I support:

- 0
- 1
- 2
- 3
- 4 of these things

How many of the following things do you personally support? You don't need to say which ones you support, just specify the number of them (0, 1, 2, or 3).

State measures to prevent abortion

Legalization of same-sex marriage in Russia

Increase in monthly allowances for low-income Russian families

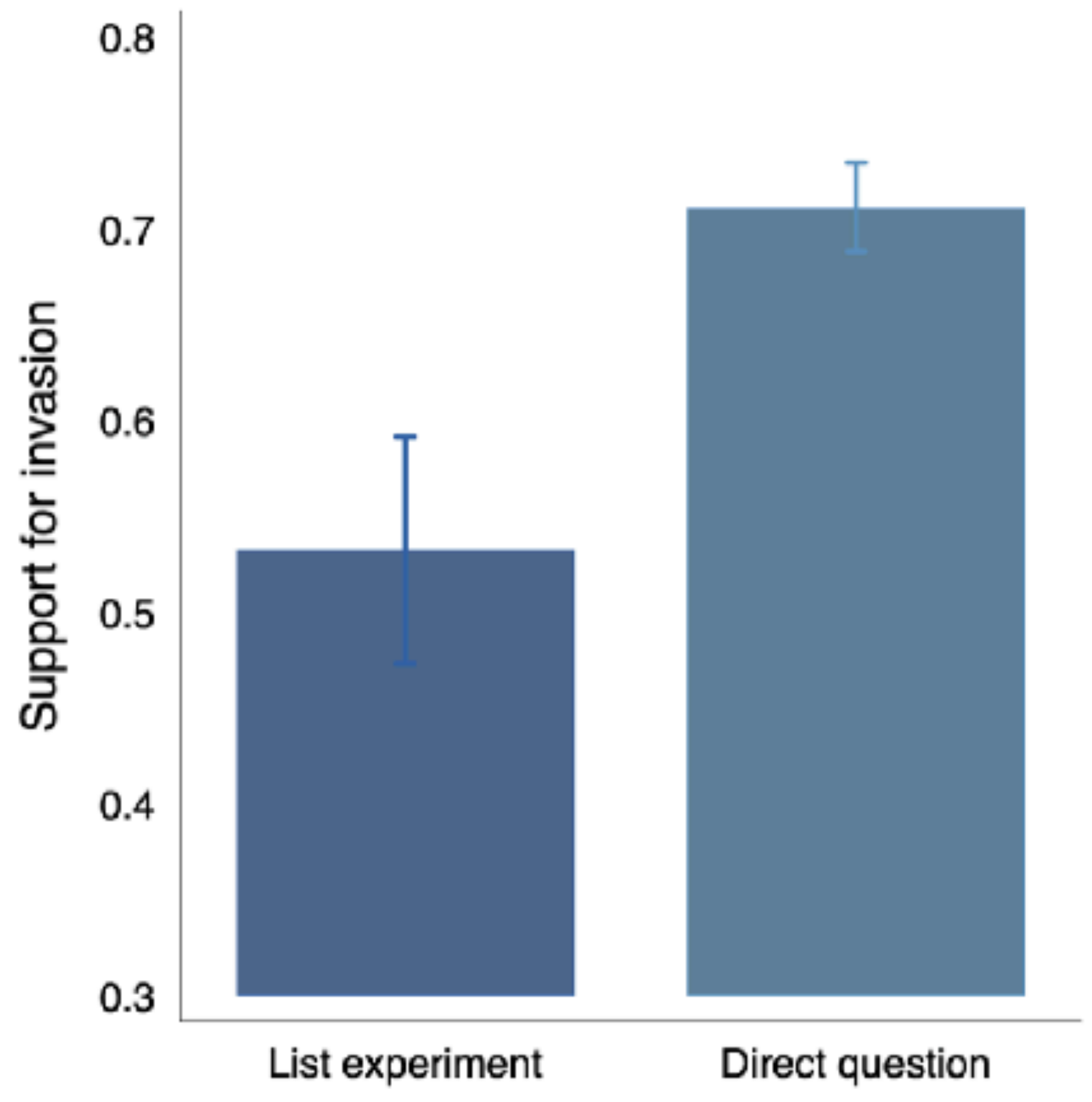
I support:

- 0
- 1
- 2
- 3 of these things

Chapkovski and Schaub 2022. "Do Russians tell the truth when they say they support the war in Ukraine? Evidence from a list experiment" LSE Blog

Aside: how is Putin’s popularity calculated?
List Experiment

Figure 2: Support for the Russian invasion of Ukraine



Note: Bars show averages, vertical lines show 95% confidence intervals.

Chapkovski and Schaub 2022. “Do Russians tell the truth when they say they support the war in Ukraine? Evidence from a list experiment” LSE Blog

Agenda

01

Logistics

02

Why Privacy

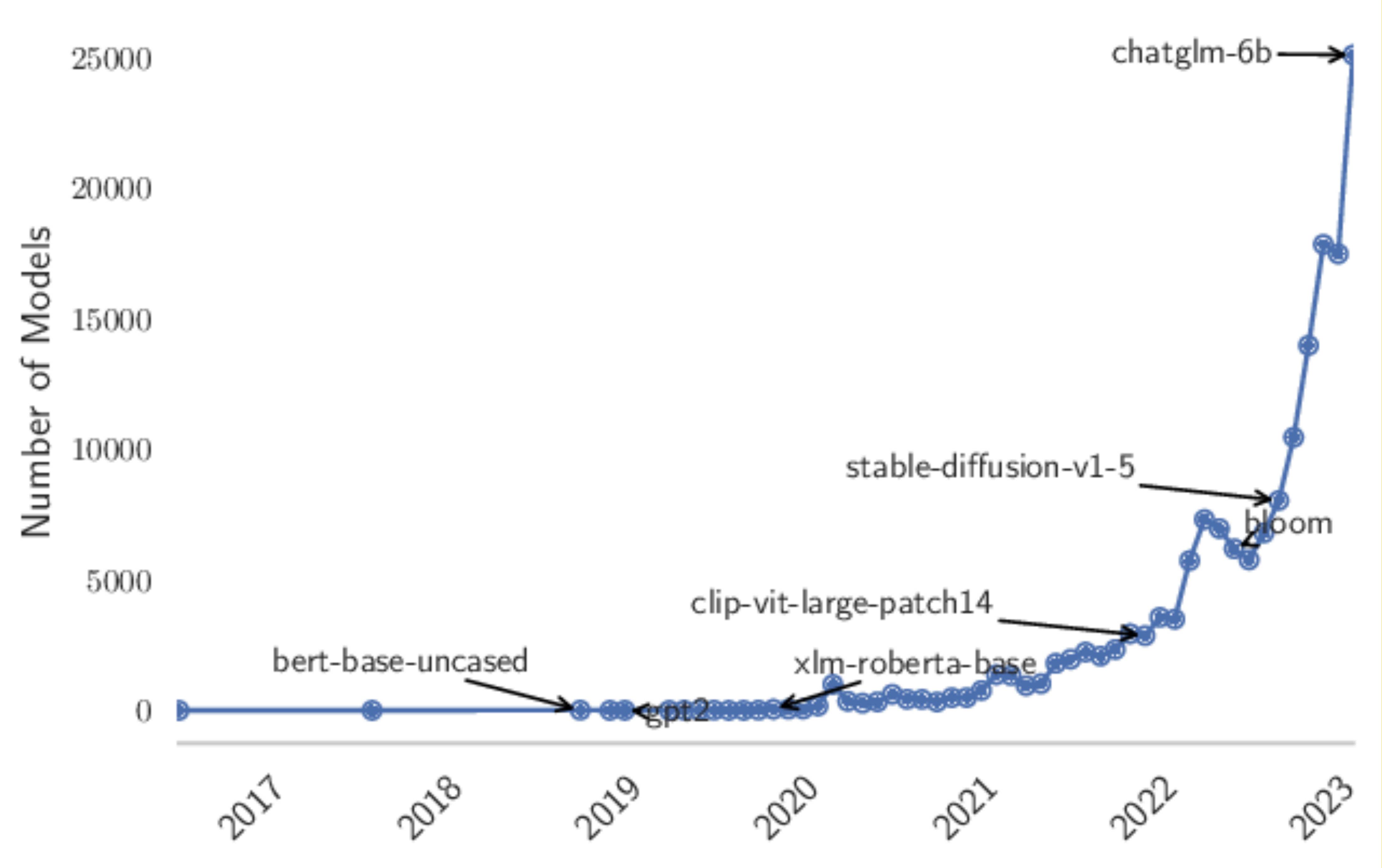
03

What Privacy

04

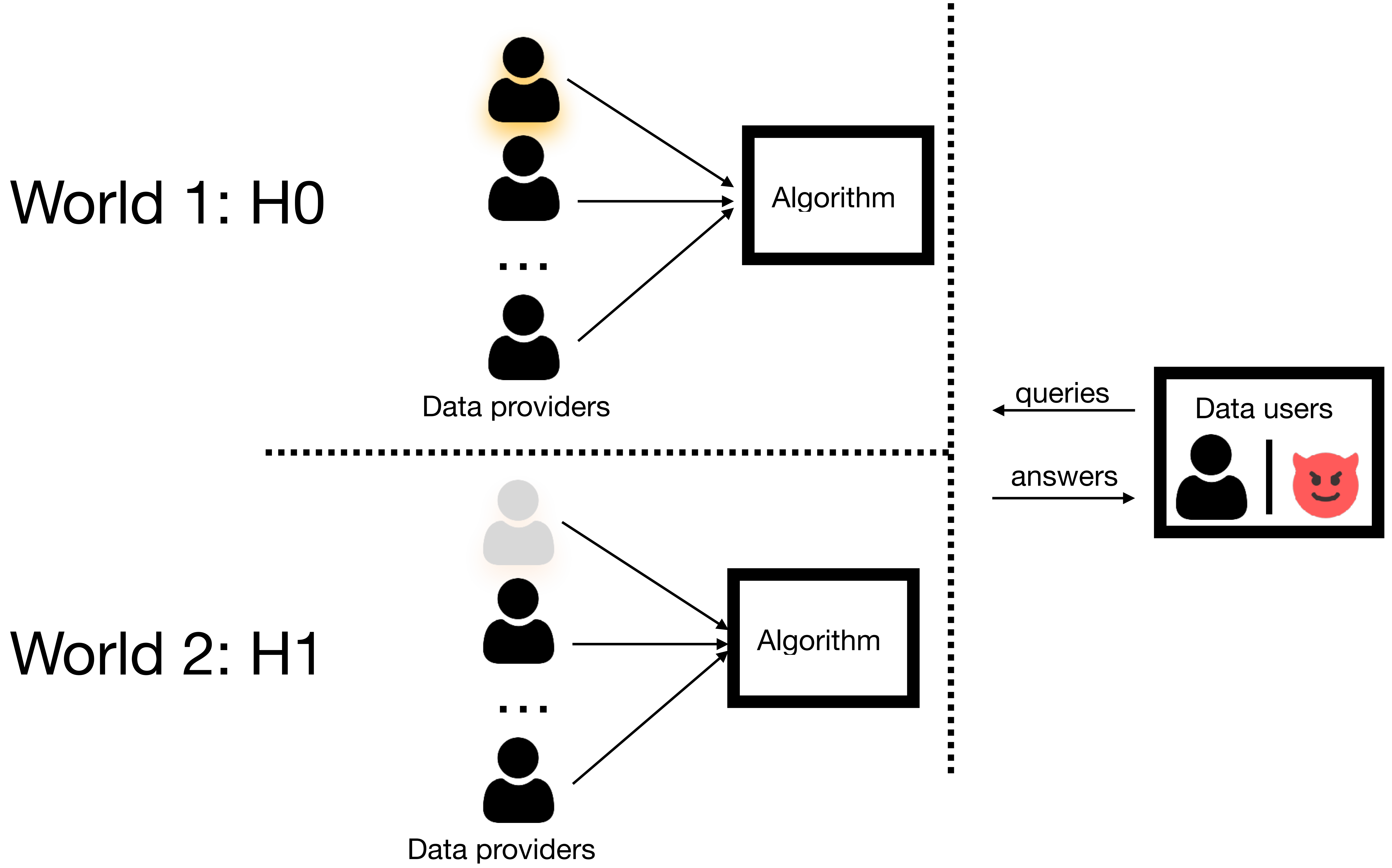
Privacy in ML

Lots of models being released



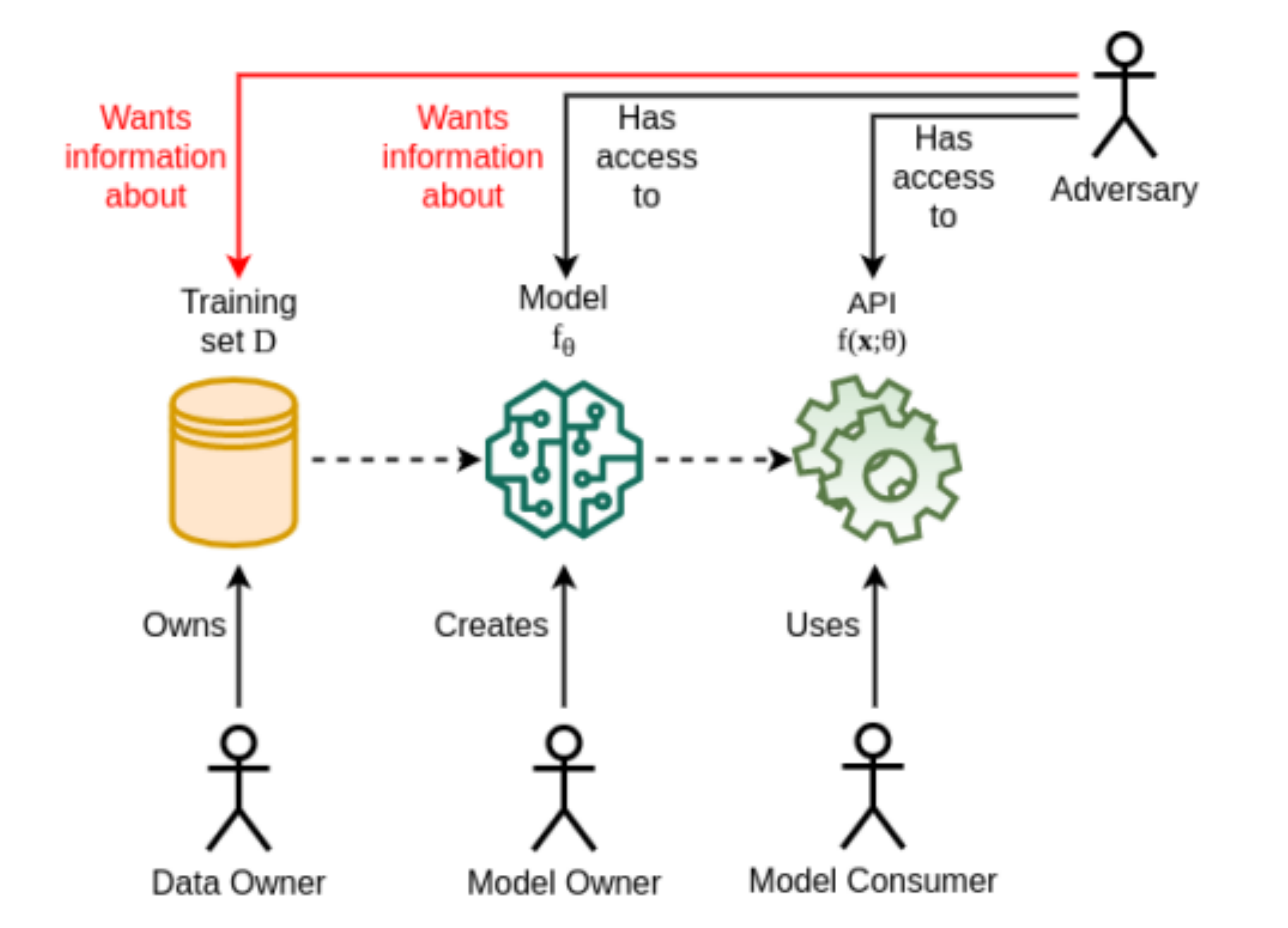
Quantifying Privacy Leakage

Attempt 2



- In world 2 only Bob is removed/ replaced.
- Now from the answer, how easily can guess the correct world?

ML attack taxonomy



Threat model [Cristofaro 2020]

Kinds of privacy attacks in ML

- White-box vs black-box: what level of access do you have?
- Training time vs. test time attacks: when does the attack take place?
- Active vs. passive: how much influence do you have?
- What do you want to steal?
  - model architecture?
  - model parameters?
  - reconstruct training data?
  - infer attribute of a datapoint?

Extracting data from ML models

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

xkcd 2169

Extracting data from ML models

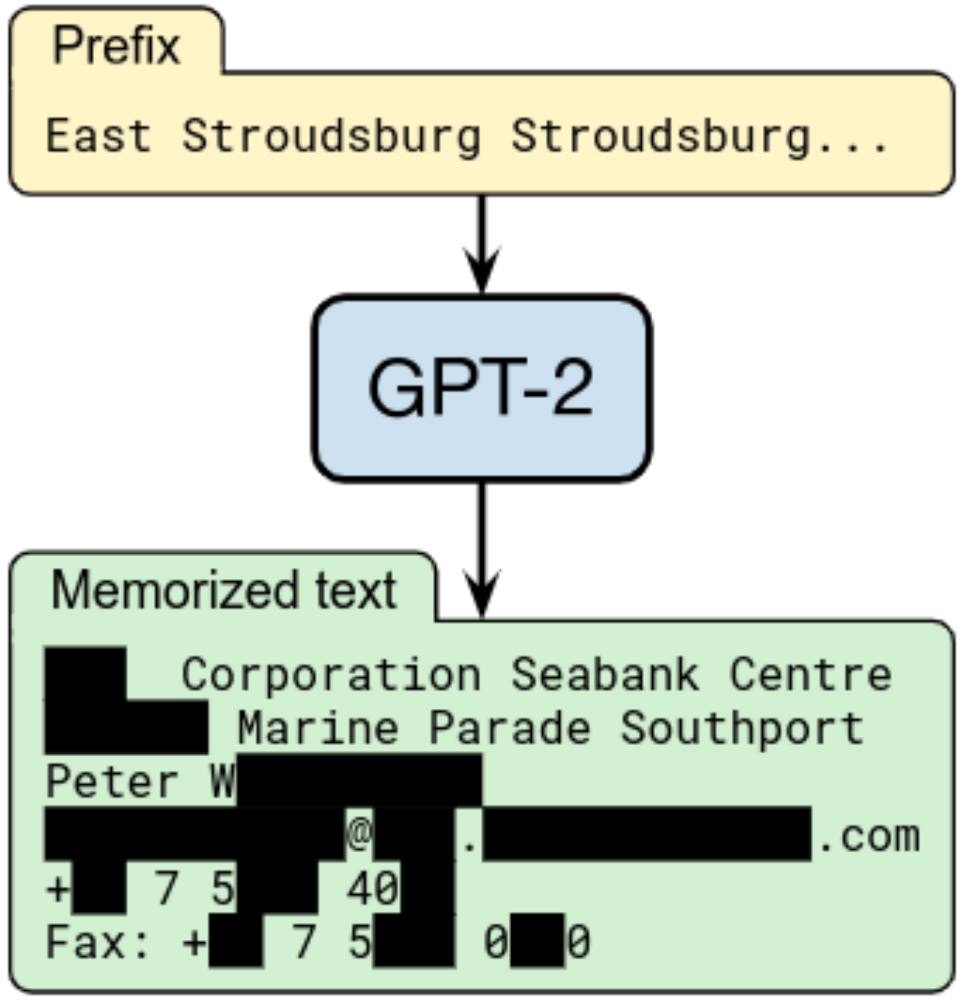


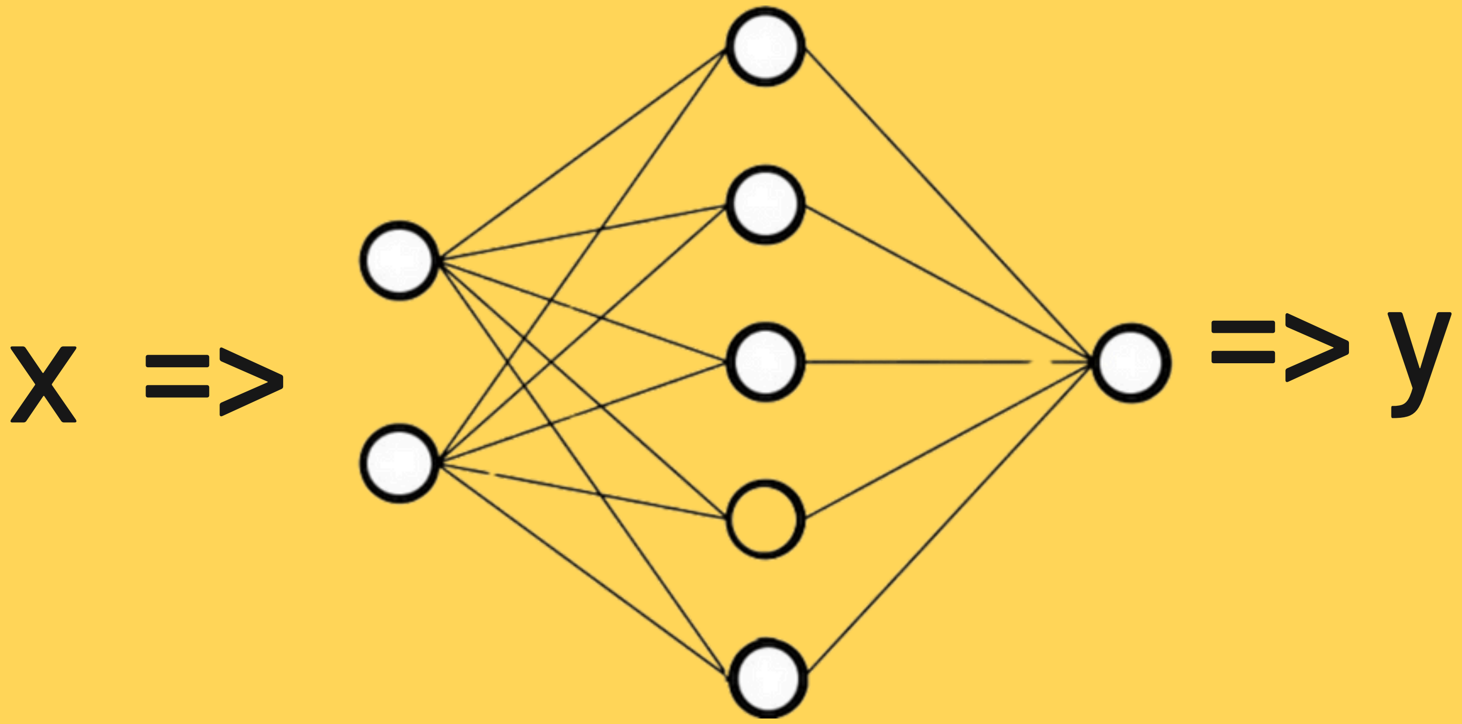
Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]

[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

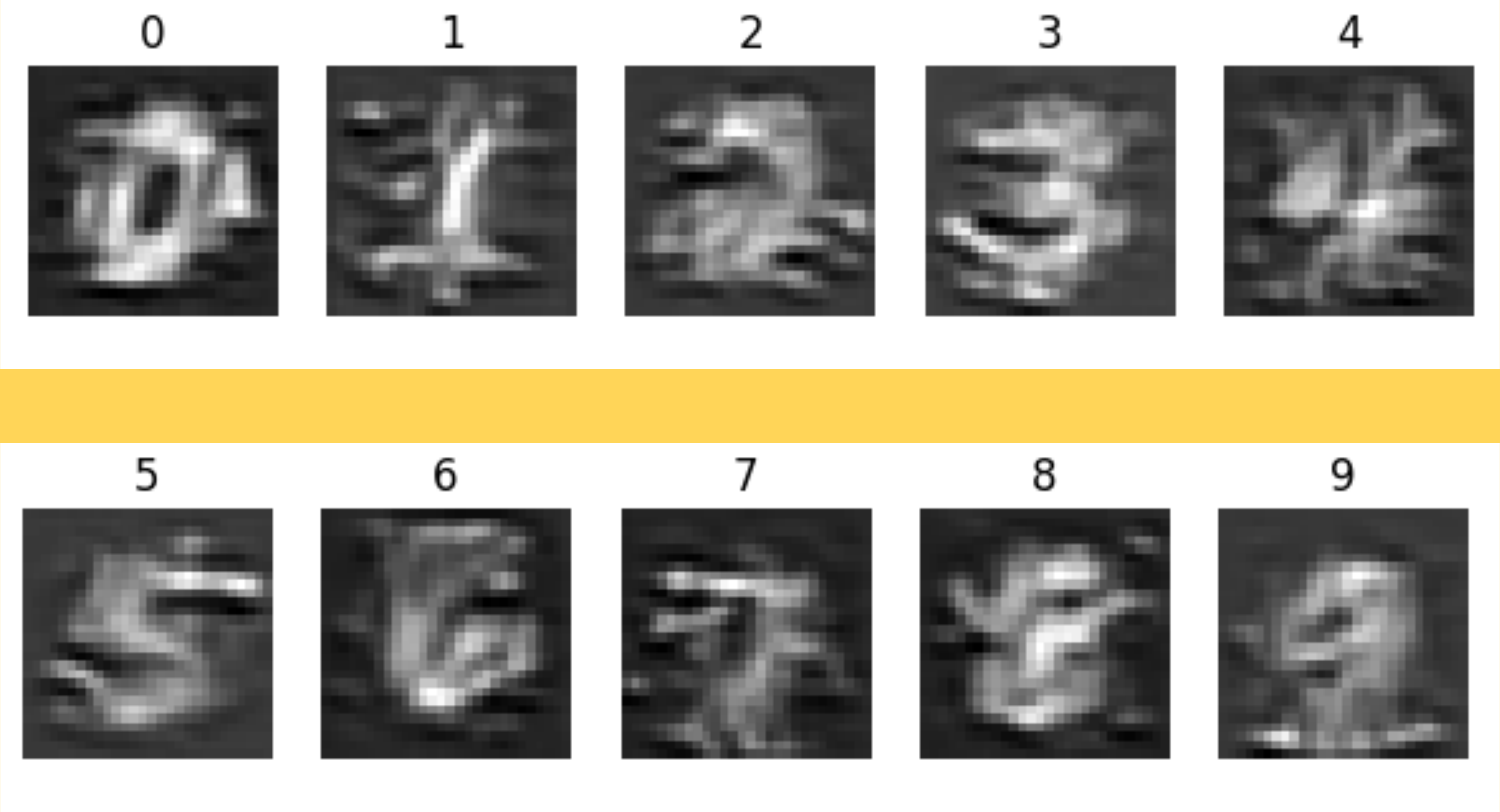
Model inversion



- Idea: model will be more confident on an image it has seen in training
- optimize over ***x*** such that ***y_label*** is high.

$$\min_{x} \ell(f(x), y)$$

Model inversion



See jupyter notebook.