CSCI 699: Privacy-Preserving Machine Learning

Sai Praneeth Karimireddy

Agenda

01

Logistics

02

Why Privacy

03

What Privacy

04

Privacy in ML

Course Logisitics

- Fri 1pm to 4:20pm.
- Room currently ~~DMC 200. **Might change** - keep an eye out!~~ WPH 102
- Course website: spkreddy.org/ppmlfall2024.html
- Email: karimire@usc.edu (add CSCI 699 in subject)
- Anonymous feedback: https://forms.gle/EqRmkZhgMrtgDh2o9

Course overview

- What even is privacy?
- How to know if my data was used to train ChatGPT?
- How can you train a model while guaranteeing the privacy of the data?
- 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?

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.

Grading

- 3 Assignments: **30**%
  - short: checking your understanding of the core concepts

- Presentations: **25**%
  - 20mins presentation
  - Pick 1 paper from list - subset of additional readings.
  - Pick any date **after** the day of topic.
  - Signup sheet will be sent out eod. First come first serve. **Sign up by Oct 1.**

- Report: **30**% Due exam day, more details next.

- Discussion: **15**% more details next.

Report: 30%

Option 1

- Team up with others who signed up for the same topic - 1 to 3.
- Teach each other your papers and related background.
- Write up a 4 page report. Formatting instructions to follow.

Option 2

- Team up with 2-3 others.
- Come up with an research question (based on what you've read or otherwise)
- Setup a meeting to get my feedback **before Oct 15.**
- Write up a 4 page report.

Discussion: 15%

Before start of class of presentation, submit 1 paragraph per paper being presented from any of the following perspectives:

- **Reviewer #3**: One really good reason and one bad reason why this paper should have been rejected. NeurIPS guidelines
- **Industry practitioner**: how will you make a great product out of this paper?
- **Researcher**: abstract of an impactful follow-up work

Agenda

01

Logistics

02

Why Privacy

03

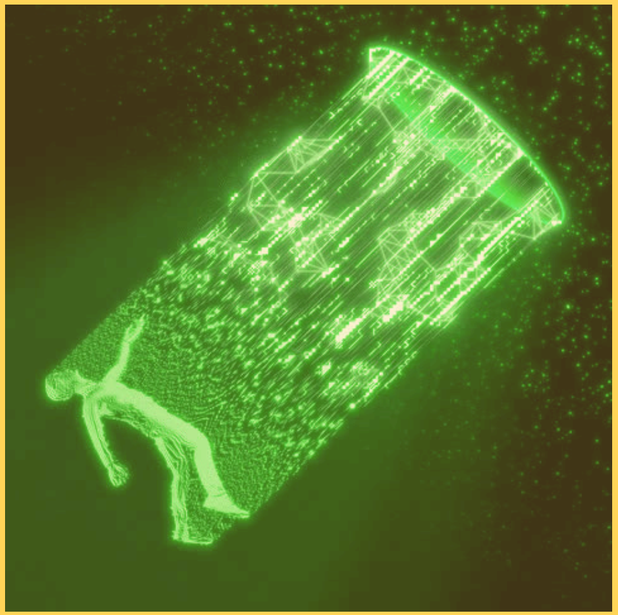
What Privacy

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

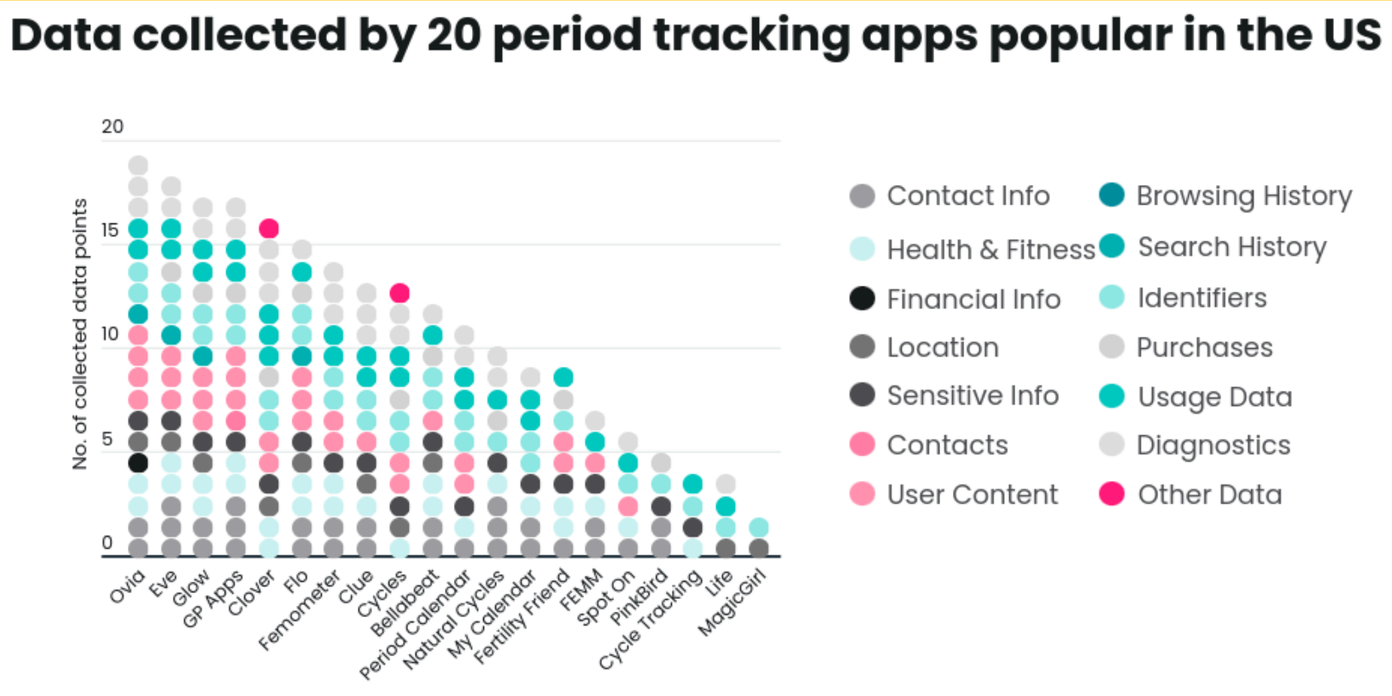
2010

“No one cares”

“If you’ve done nothing wrong…”

2009

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

FBI wanted poster, CBC News 2018.

- You may be very careful. But doesn't matter.
- 23AndMe has genetic information of 15 million people.
- National DNA Index (NDIS) contains about the same, but only of offenders.

Why privacy? Case 3

FBI wanted poster, CBC News 2018.

- Bayes time!
- Probability of 23AndMe being involved in a crime is 1%
- Probability of NDIS being involved in a crime is 10%
- DNA test is True positive = 90%, False positive = 1%

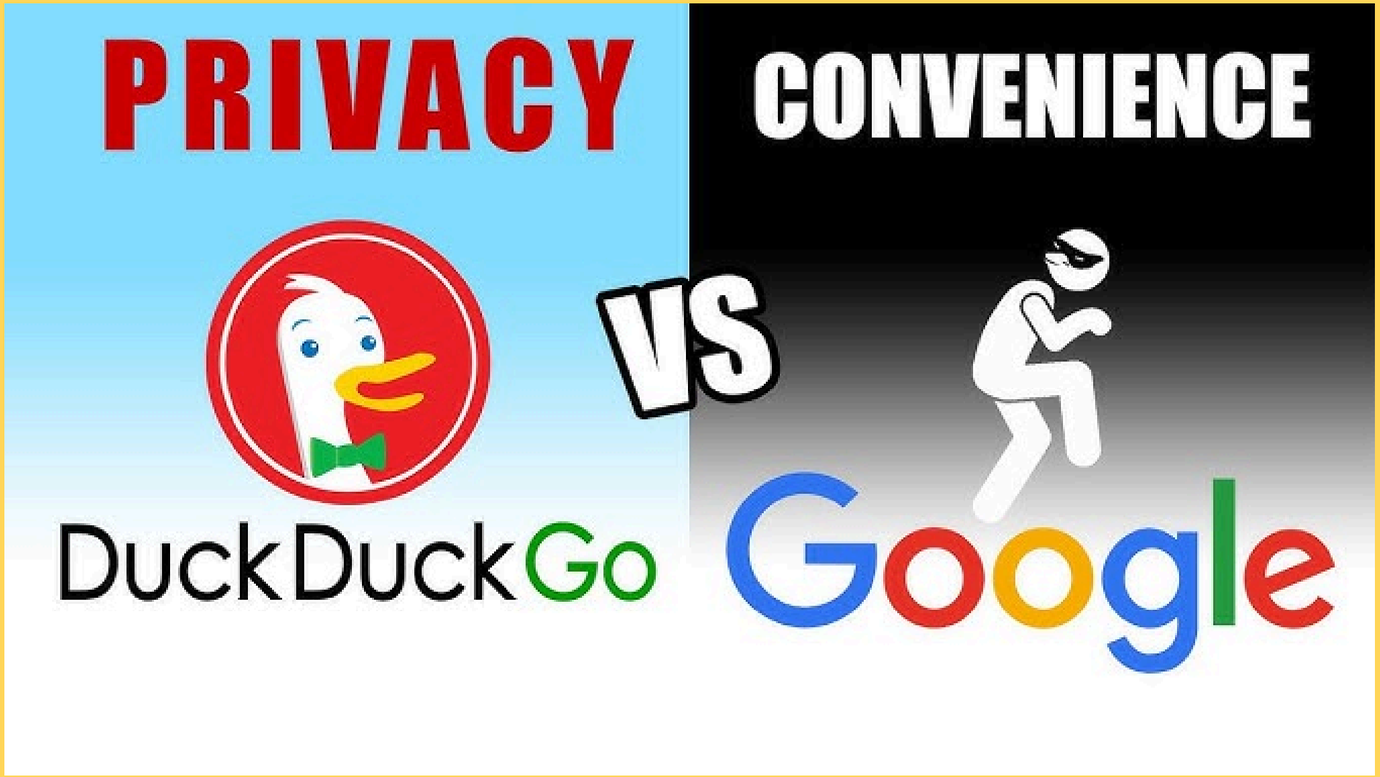
Test from NDIS said yes vs from 23AndMe?

Why privacy? Summary

EU Law analysis 2020

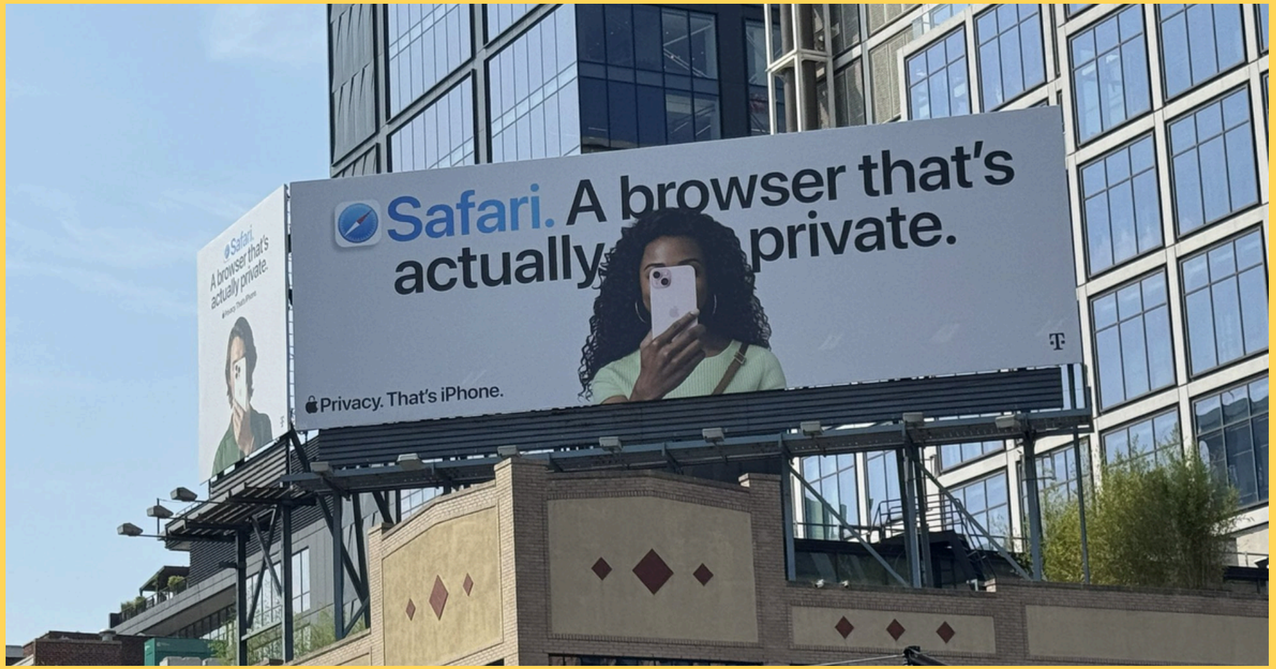
- You are being looked at, but you can’t look back.
- If a flag is raised, very expensive to deal with.
- You will change your behavior to be overly cautious and not raise flags => “chilling effect”
- Privacy is about power-imbalance.

Privacy is also BIG BUSINESS

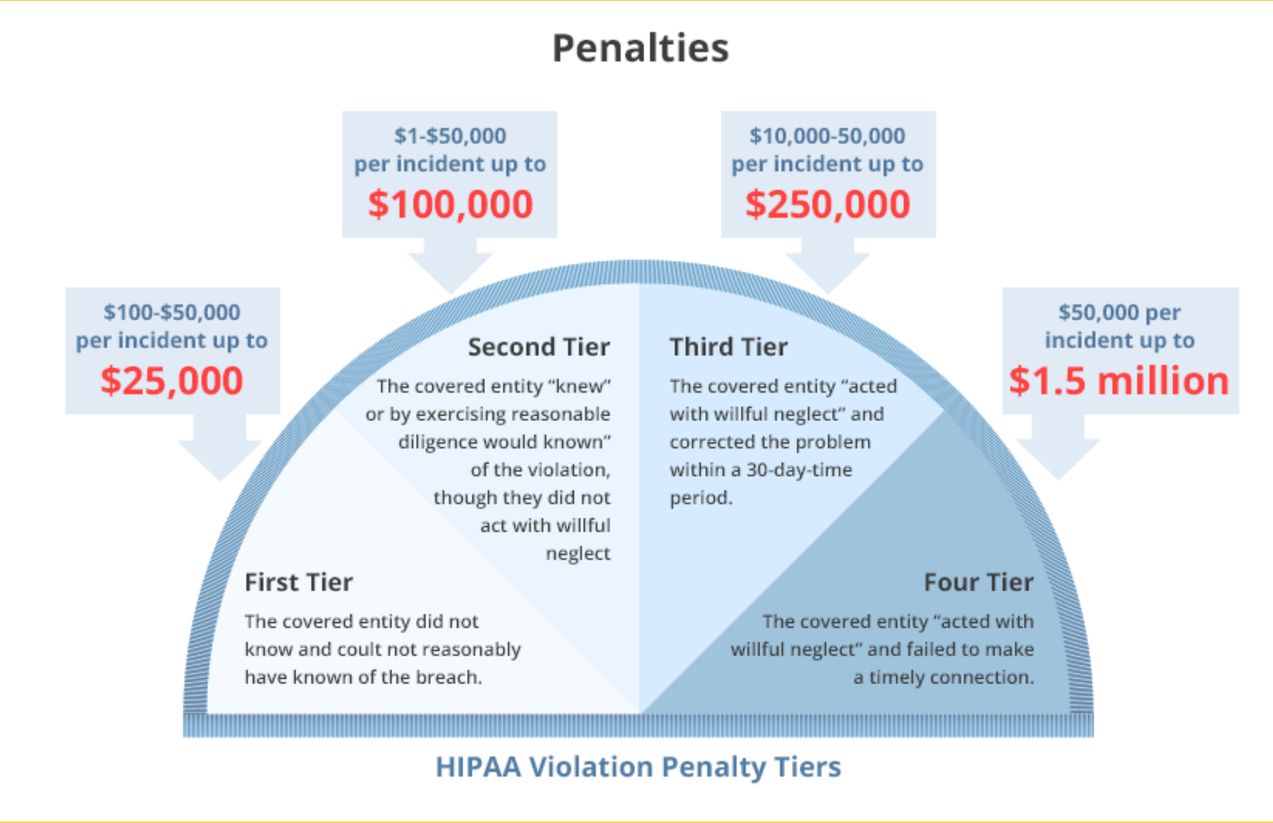


- If you don’t trust Google, you may start using alternatives
- Google will lose out!
- Lots of effort in ensuring baseline trust and privacy.

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

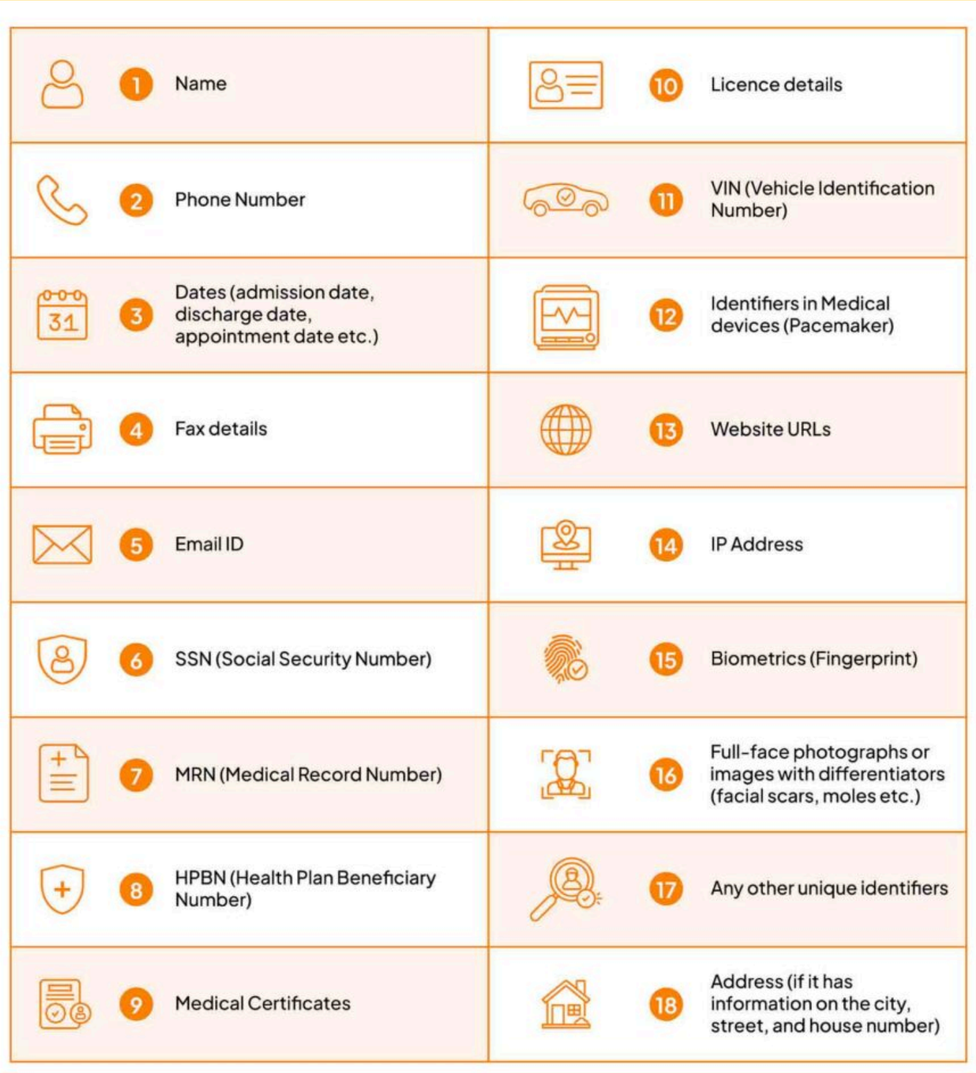
03

What Privacy

04

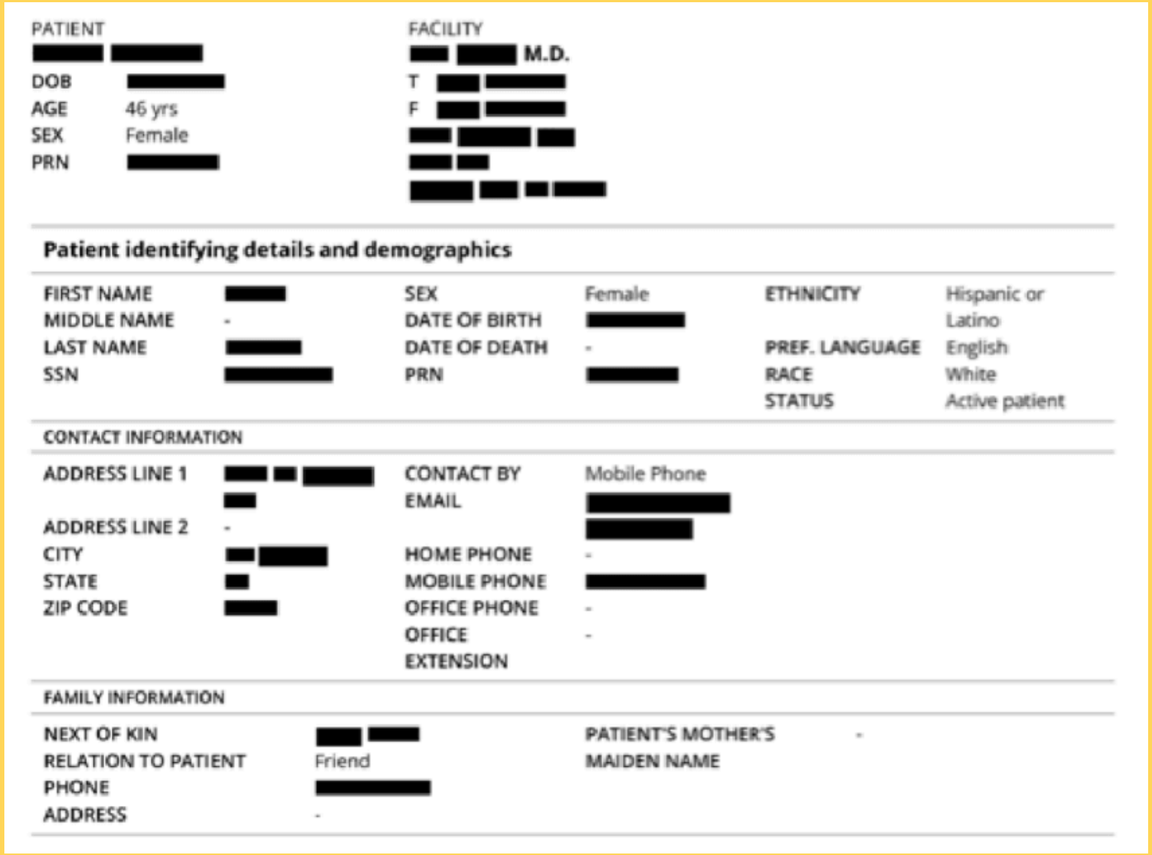
Privacy in ML

De-identification



- Remove “sensitive” and “private” attributes.
- 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	-

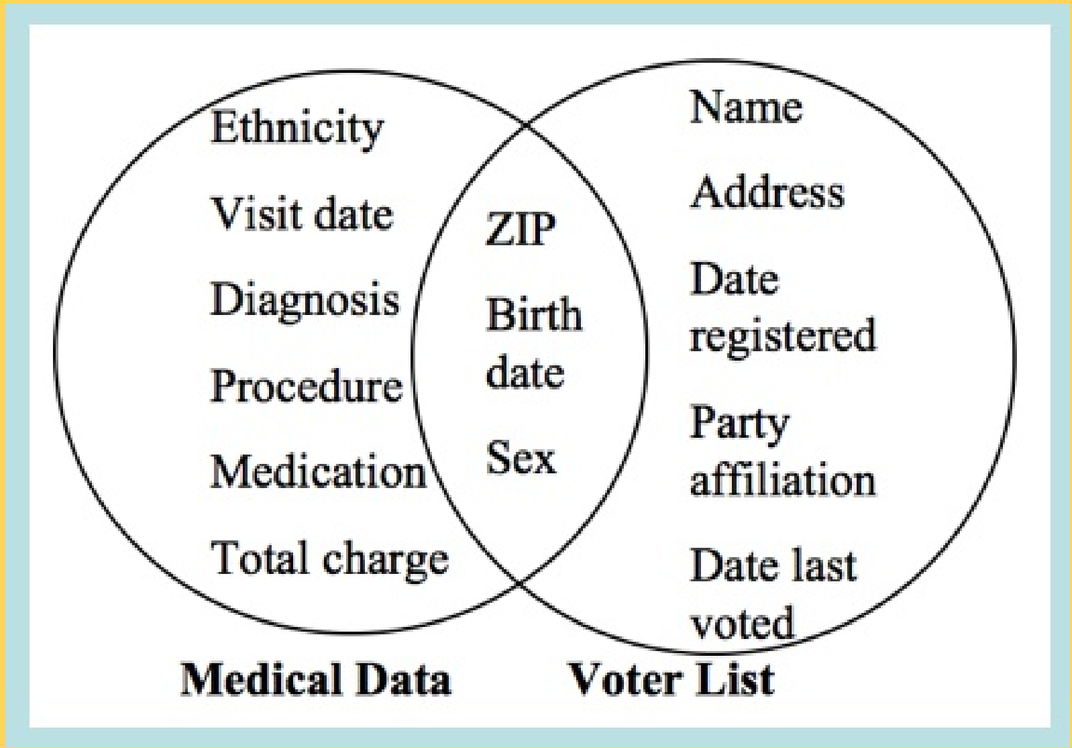
FAMILY INFORMATION

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

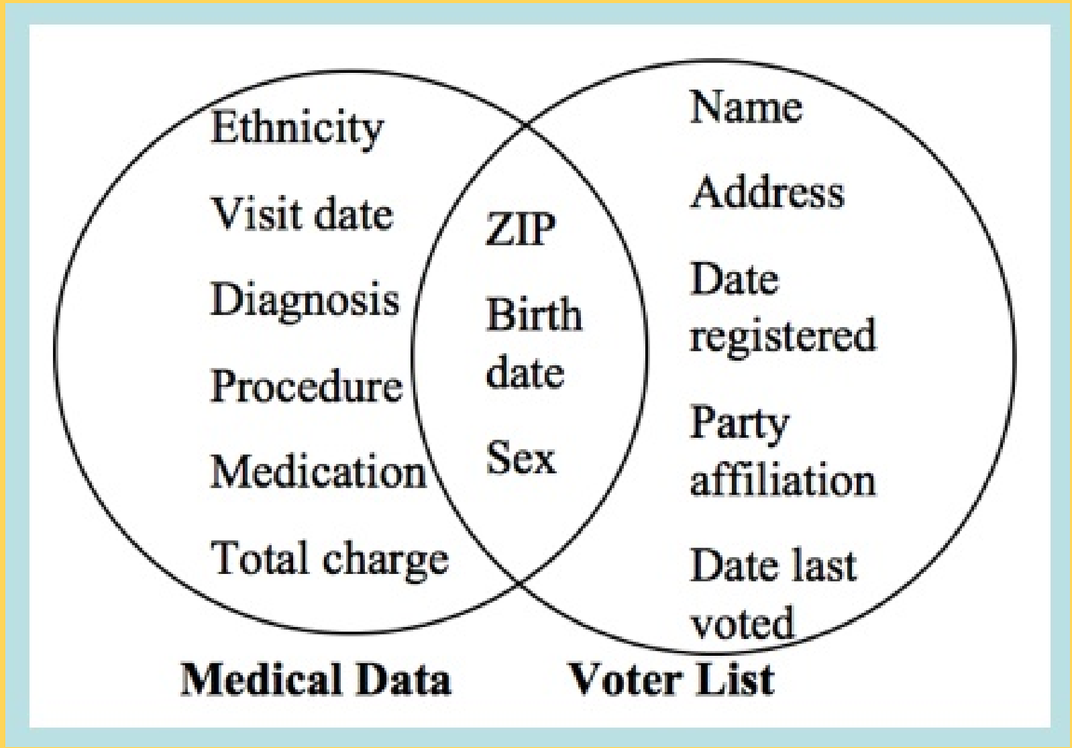
Latanya Sweeney 1997: 87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}



Bill Weld (governor): she identified his medical records and mailed them to him.

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

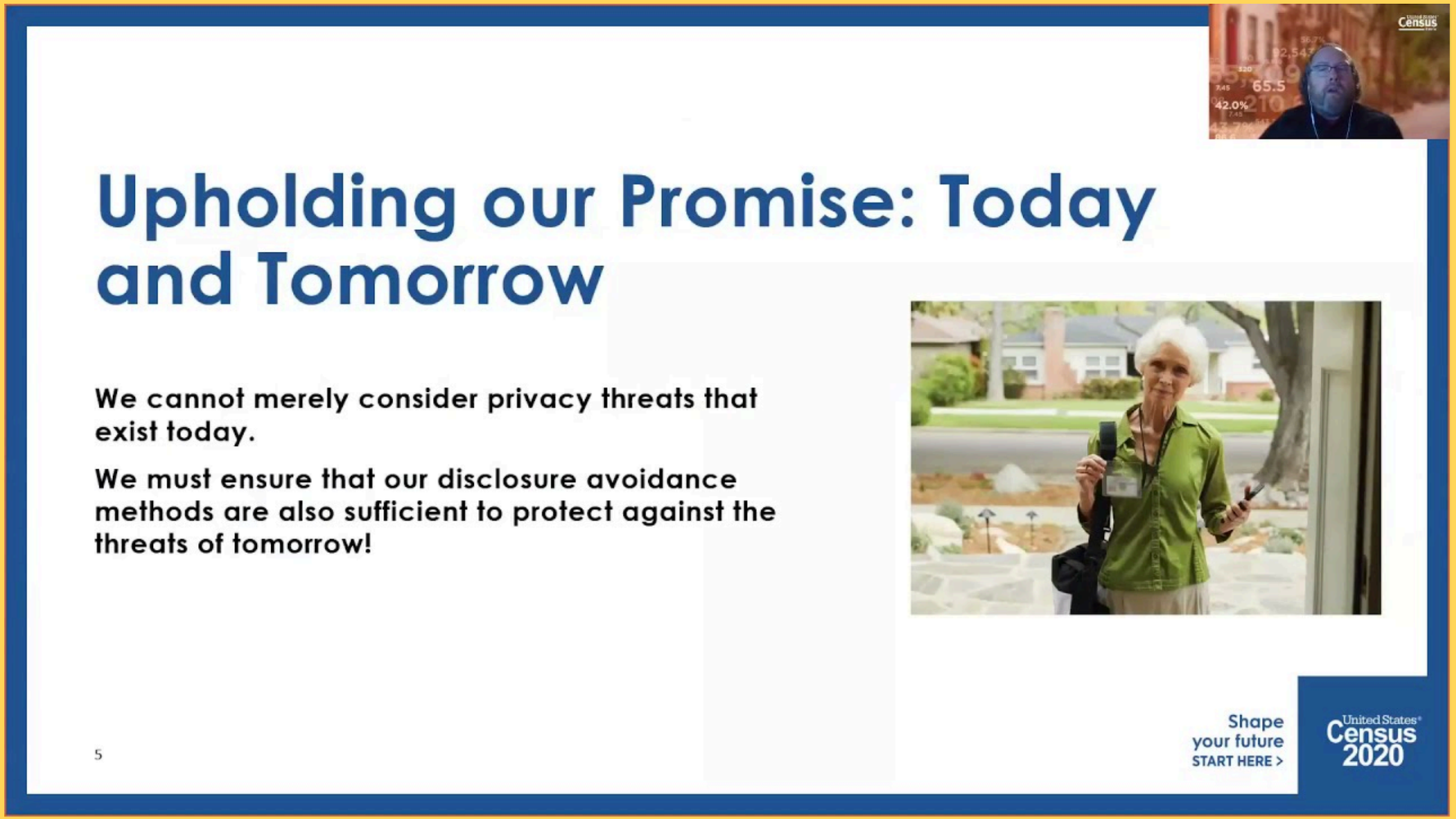
Mohammad Mahmoody
mohammad@virginia.edu

Shuang Song
shuangsong@google.com

Abhradeep Thakurta
athakurta@google.com

Florian Tramèr
tramer@cs.stanford.edu

Differential Privacy: next week



Agenda

01

Logistics

02

Why Privacy

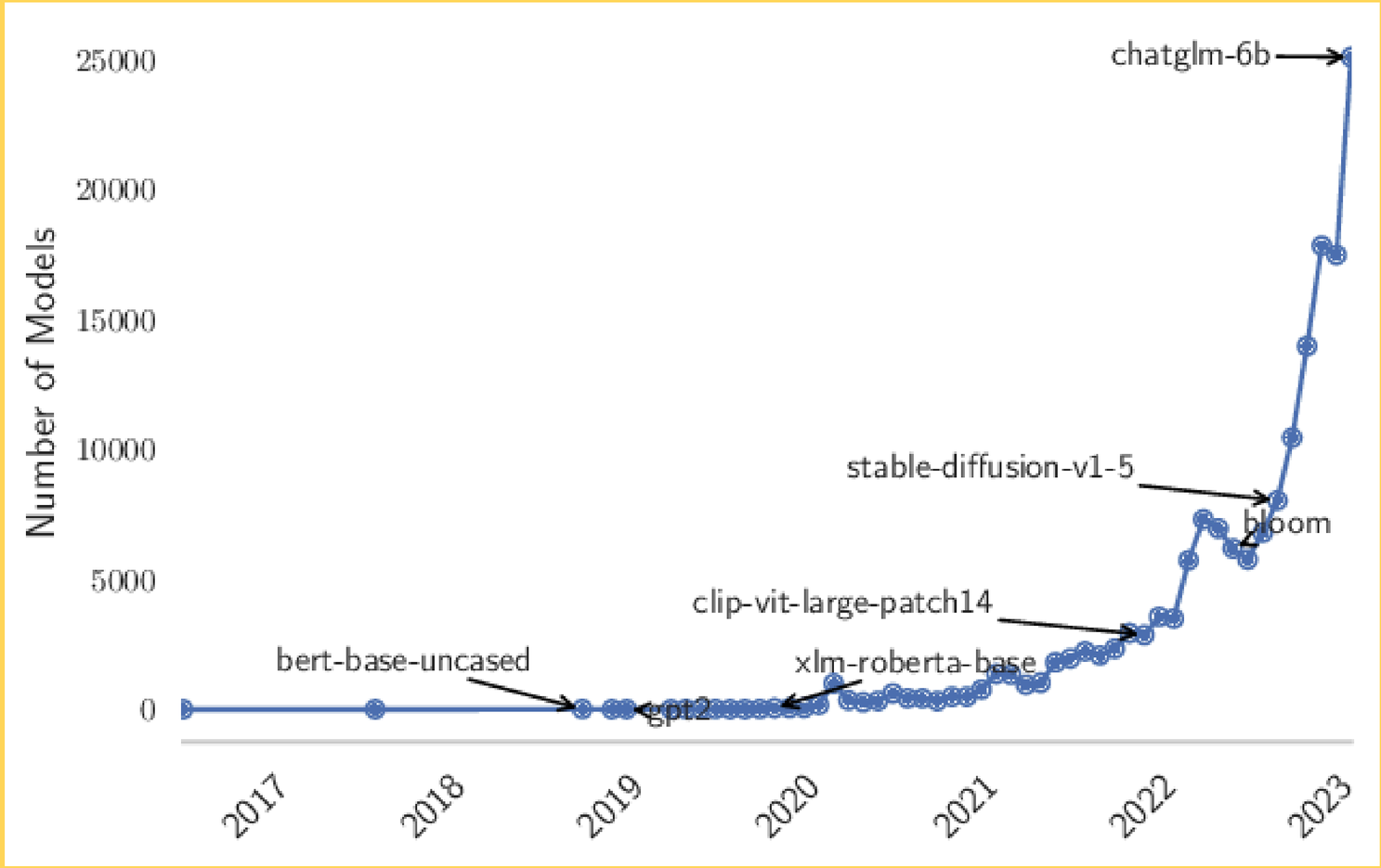
03

What Privacy

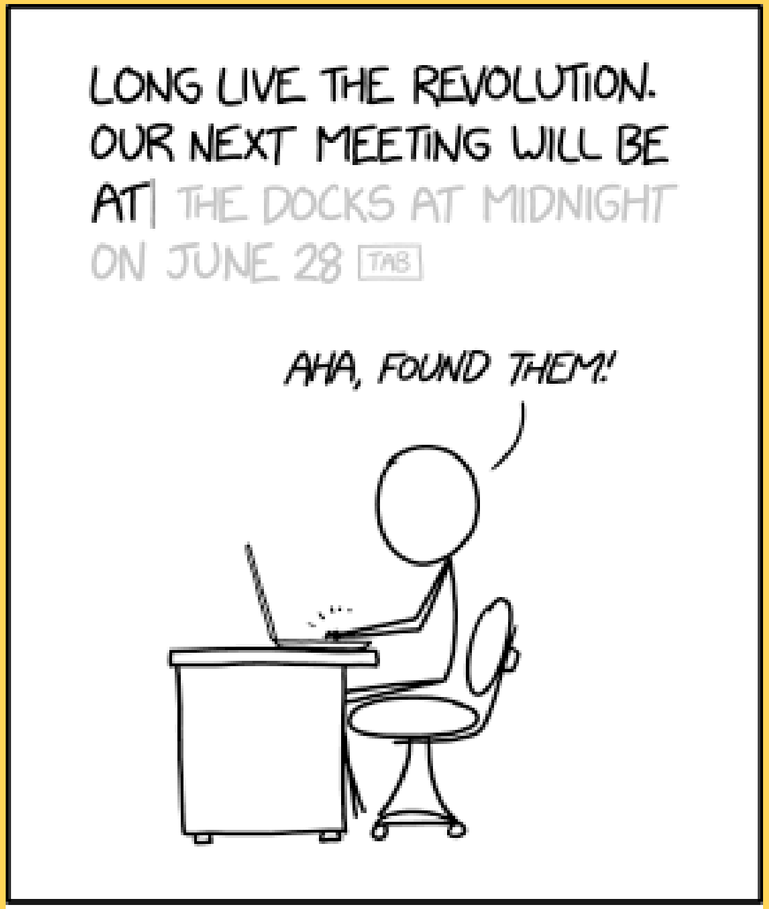
04

Privacy in ML

Lots of models being released



Extracting data from ML models



xkcd 2169

Extracting data from ML models

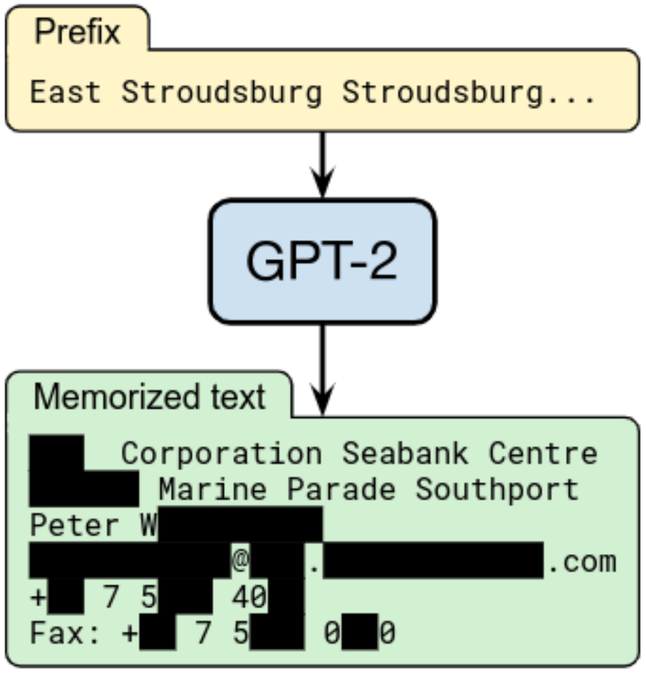
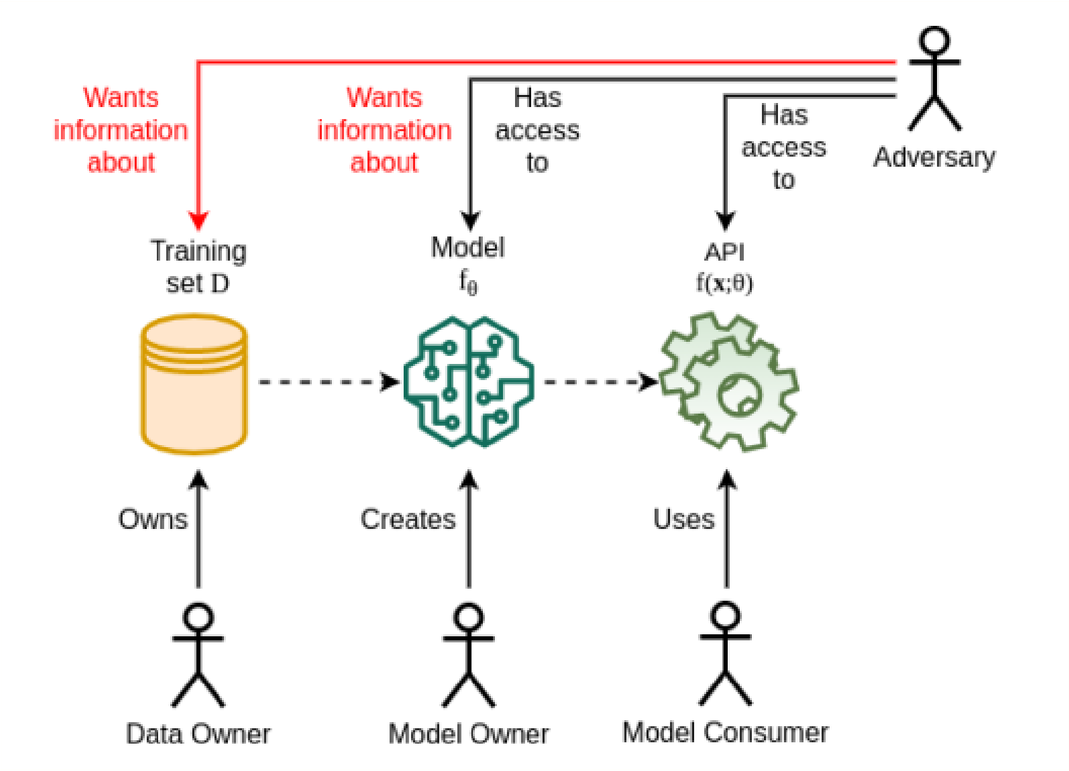


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

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*

ML attack taxonomy



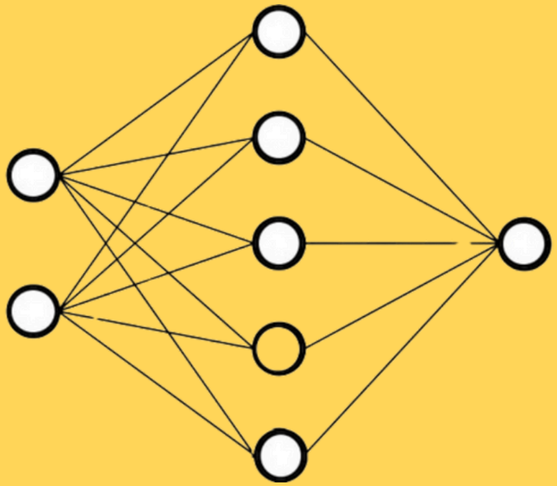
Threat model [Cristofaro 2020]

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?

Model inversion

x =>

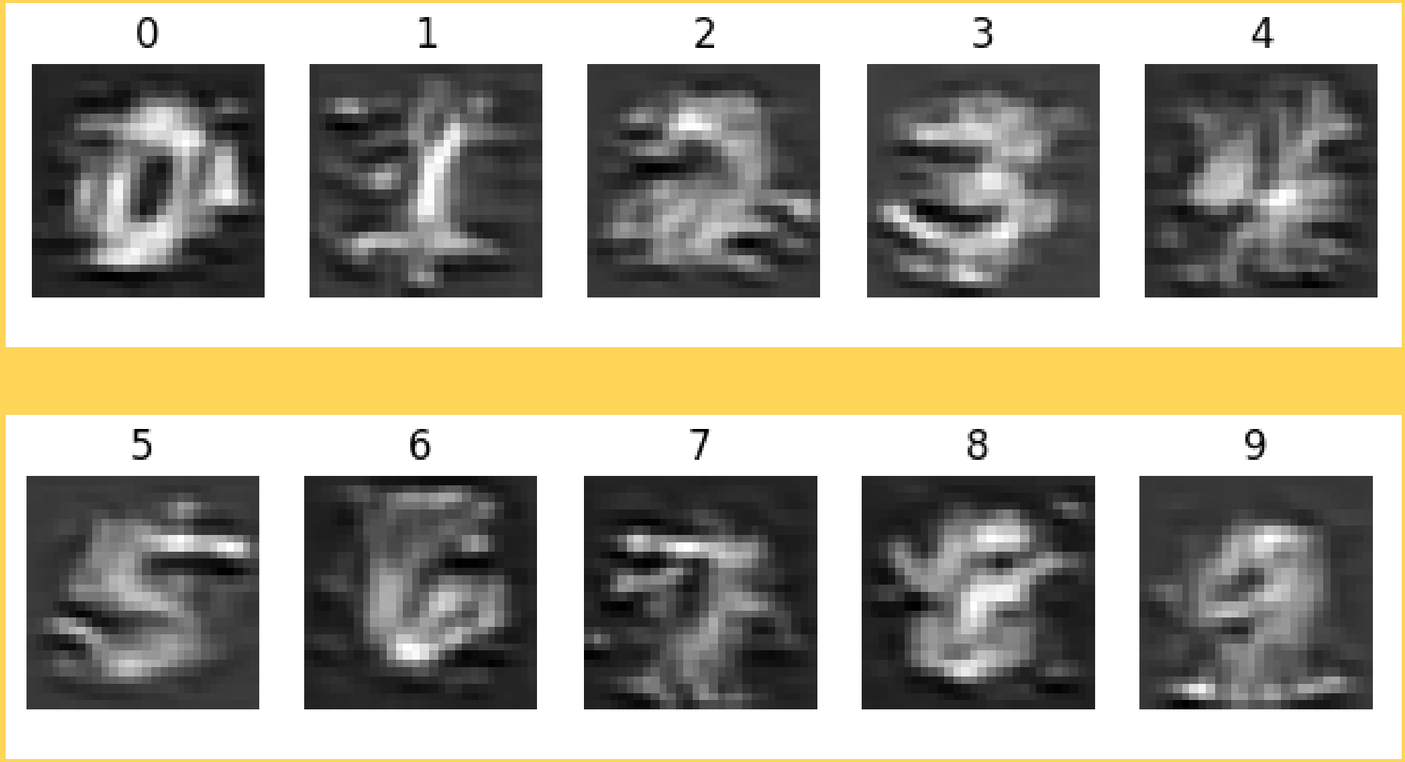


=> y

- 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.