Trustworthy Machine Learning From an Optimization Lens

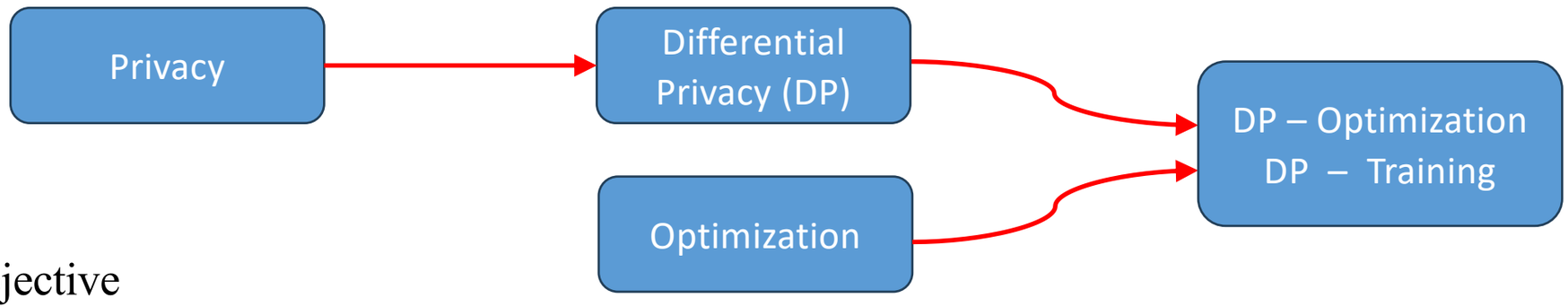
Meisam Razaviyayn

Lecture 7: Privacy Background

razaviya@usc.edu

The next few weeks on privacy

- Lectures are long
  - We should work together to keep it interactive and less boring
- Roadmap



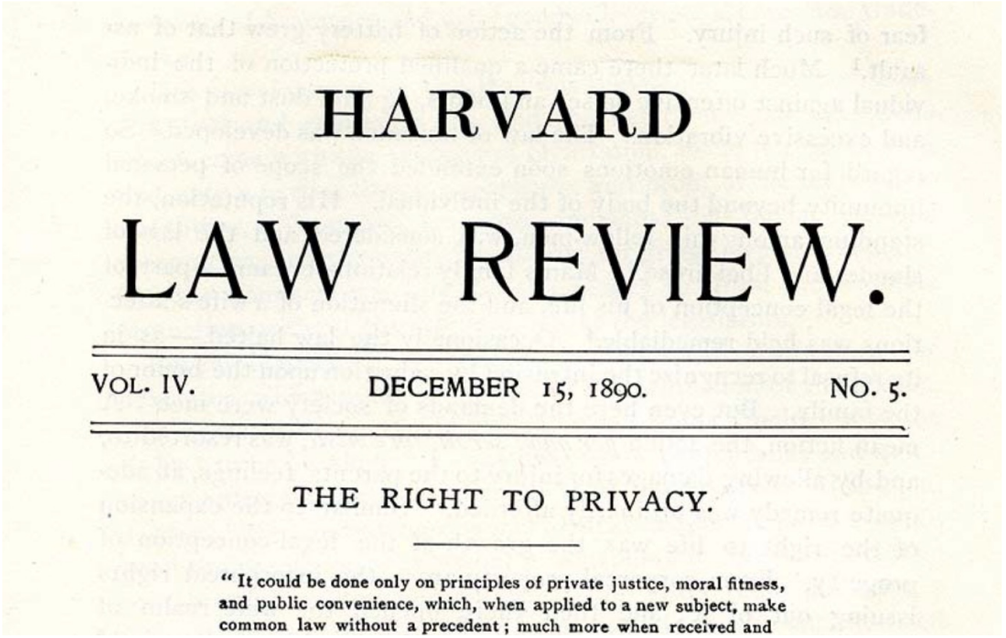
- My objective

We cover DP for the purpose of bringing it into the optimization world!

- What I want to avoid
  - Suggesting these are fully settled, well-understood concepts
  - Suggesting the solutions are already established or final
  - Suggesting that defining (differential) privacy for a specific application is straightforward
- As we go into details, keep an open mind

What is privacy?

- "The right to be let alone." (Warren and Brandeis, 1890).
  - Instant photography + yellow journalism
  - Privacy protects dignity and emotional well-being, not just secrecy or property



HARVARD

LAW REVIEW.

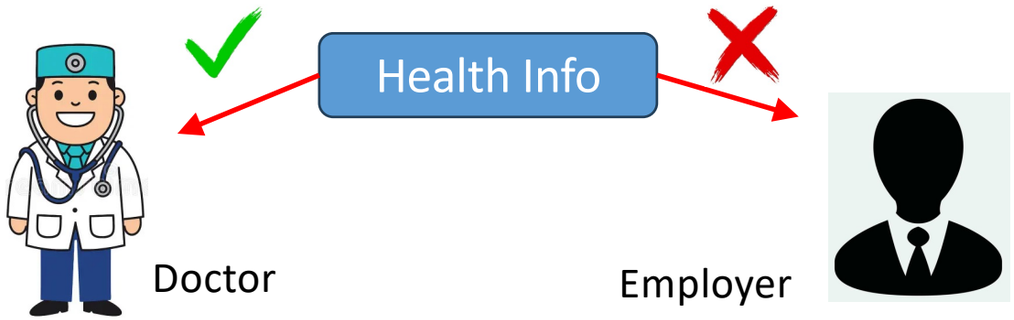
VOL. IV. DECEMBER 15, 1890. NO. 5.

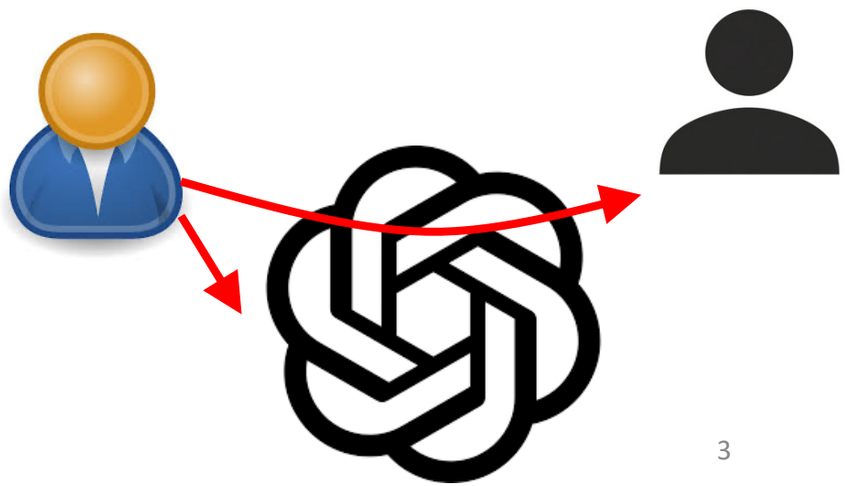
THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make common law without a precedent; much more when received and

- Modern view → appropriate use, not just secrecy

- Contextual Integrity (Helen Nissenbaum): **privacy ≈ appropriate info flow**

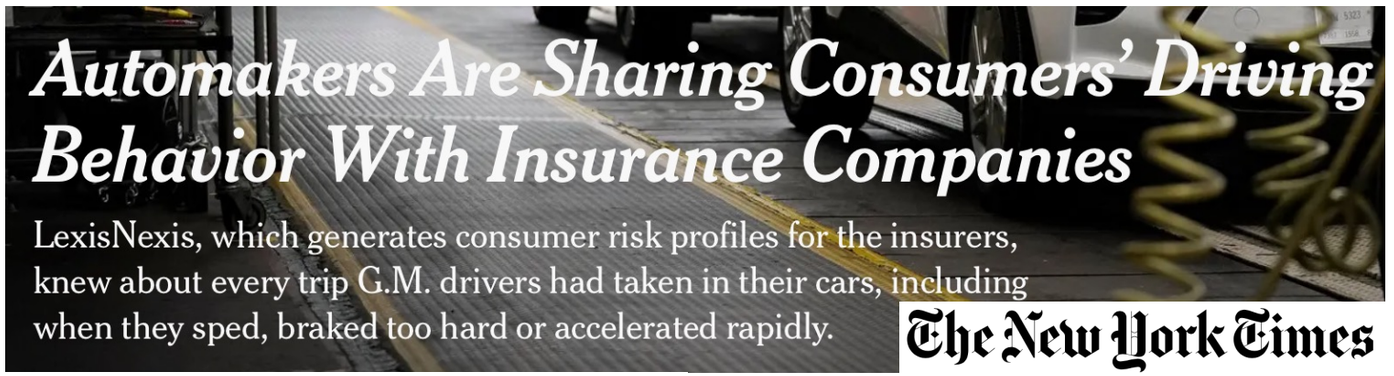




- AI often disrupts these norms.

Why privacy matters?

- We have data from credit and health data to "likes"
- **Data may reveal**: Religion, political opinions, sexual preferences, etc



Published March 11, 2024 Updated March 13, 2024

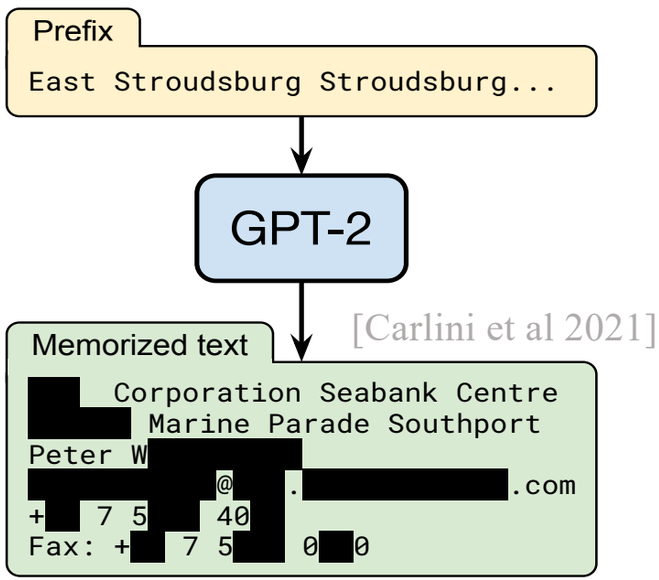
- **Potential harm**:
  - Increased insurance premium
  - Blackmailing to violent threats
  - Manipulation (e.g., Cambridge Analytica)
  - Surveillance

- Concern of many individuals

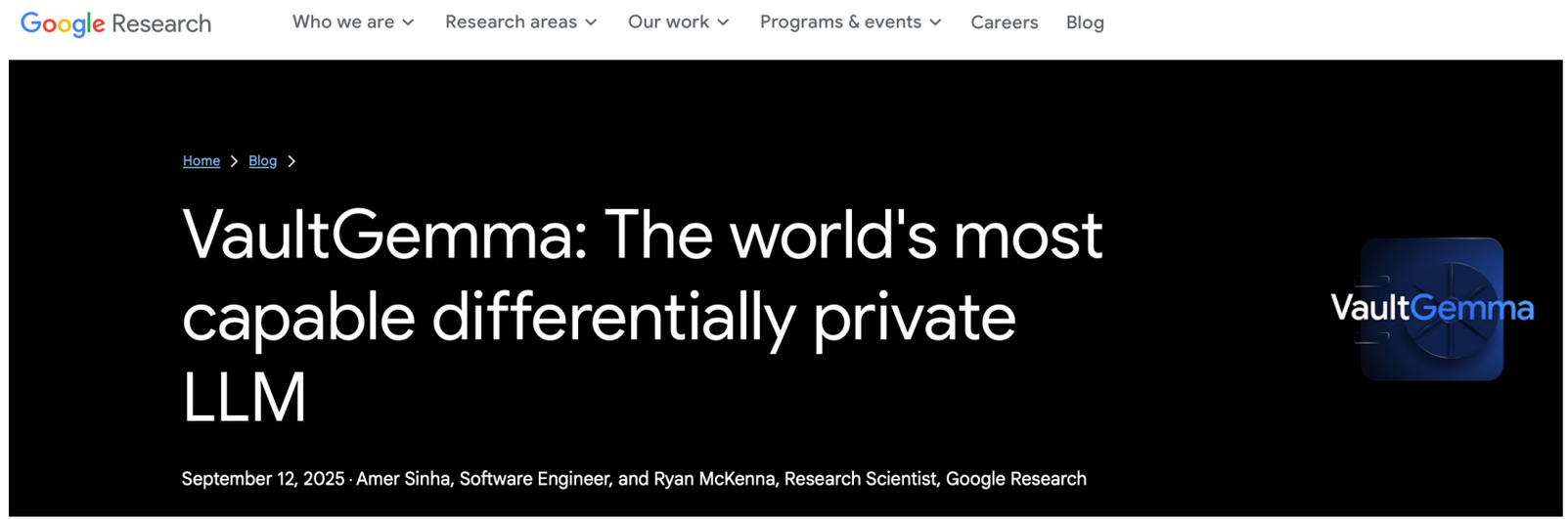
https://info.kpmg.us/news-perspectives/industry-insights-research/data-privacy-survey-orson-lucas.html
https://velocitize.com/2022/06/24/right-to-privacy-data-overload-consumer-protections/

The Age of AGI and LLMs: A New Scale of Risk

- **Data Hunger**: Foundation models train on massive datasets (e.g., Common Crawl), widening the scope of collection & exposure.
- **Inference Power :** LLMs can predict sensitive traits from seemingly harmless data.
- **Memorization**: Models act like lossy compressors, sometimes leaking snippets of training data.
- **Jailbreaking**: Safety filters can be bypassed, forcing models to reveal memorized personal information.

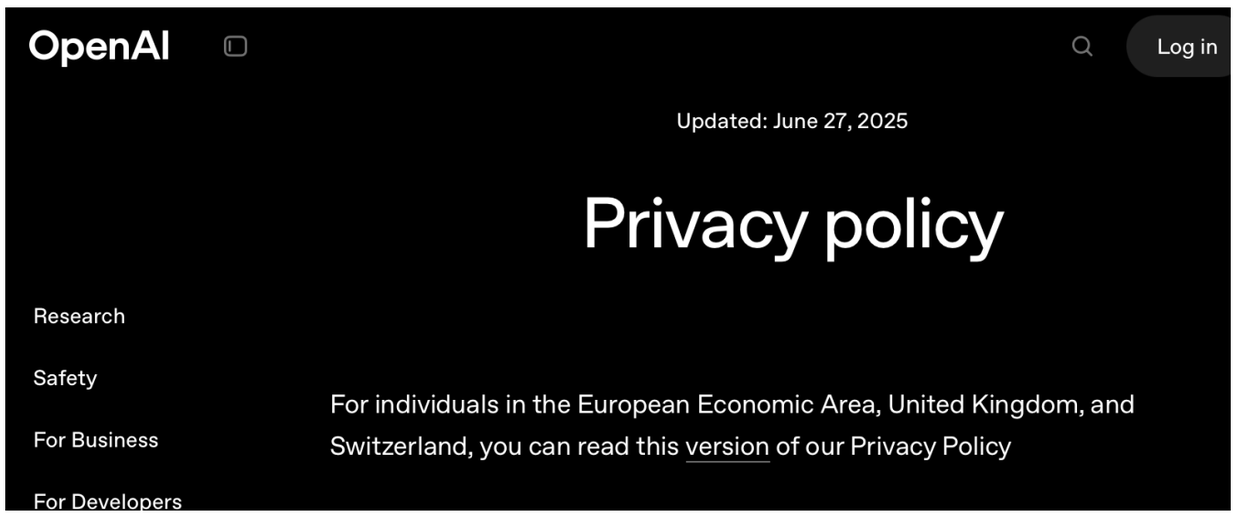


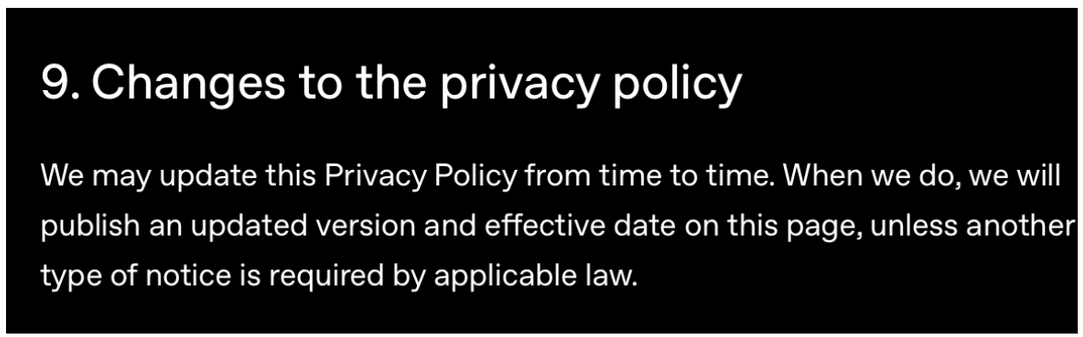
[Carlini et al 2021]



Privacy in the Headlines (2024-2025)

- **Healthcare Breaches** (2024): Change Healthcare attack exposed vulnerabilities in health data systems
- **Data Broker Scrutiny** (2024–2025) : FTC penalized companies (e.g., Avast, X-Mode) for selling sensitive location and browsing data
  - What may location data reveal?
- **LLMs and proprietary data**: Samsung employees leaked code via ChatGPT
- Privacy policy

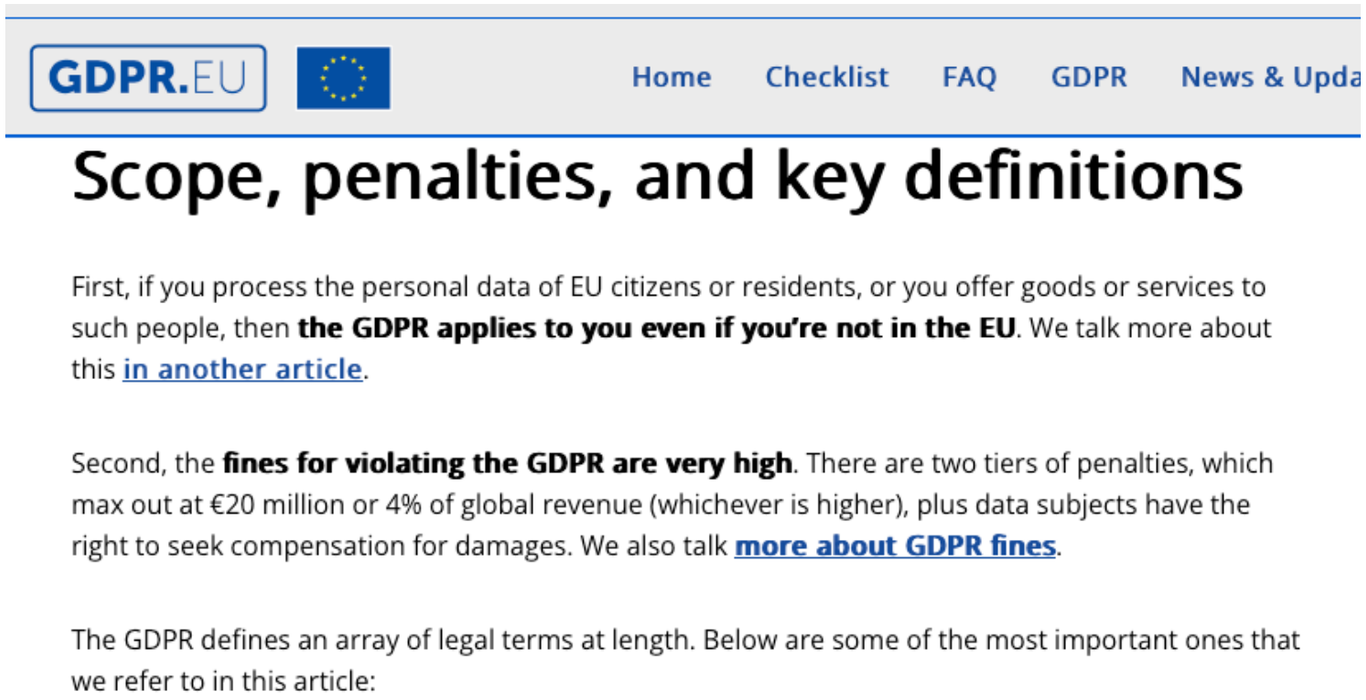




Regulations on Privacy

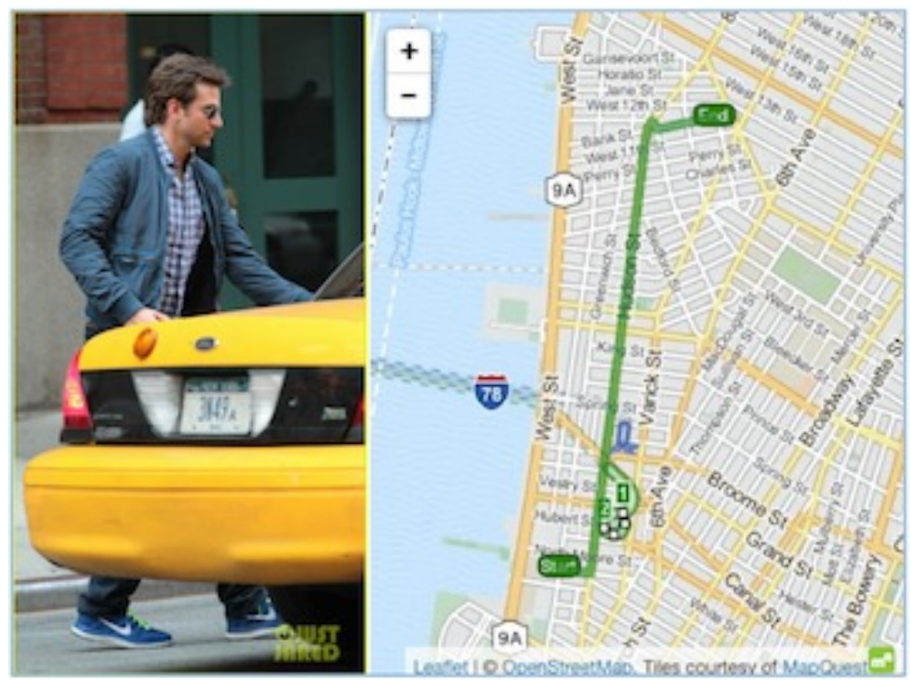
- European Union's General Data Protection Regulation (GDPR)
- California Consumer Privacy Act of 2018
- AI Bill of Rights

- Key Principles:
  - Privacy by Design
  - Data minimization
- Right to be forgotten



The Illusion of Anonymization (PII Removal)

- **Naive Approach**: Simply removing obvious Personally Identifiable Information (PII) like names or SSNs
- **Quasi-Identifiers (QIs)**: Attributes that don't uniquely identify on their own but can re-identify individuals when combined (e.g., Zipcode, Age, Gender).
- **Real-World Example 1**: NYC Taxi Data
  - Originally anonymized by removing driver names and medallion numbers
  - Researchers combined timestamps, pickup/dropoff locations (QIs) to re-identify drivers or trips
  - Reddit discussion: https://www.reddit.com/r/bigquery/comments/28ialf/comment/cicr3n2/
  - Chris Whong Analysis: https://chriswhong.com/open-data/foil_nyc_taxi/



Anonymization ≠ Privacy

Real Example 2: Governor Weld

- **Background**:
  - In the early 1990s, the Massachusetts Group Insurance Commission (GIC) released "anonymized" health data for research purposes.
  - The dataset contained patient medical records, but names, addresses, and other direct identifiers were removed.
- Latanya Sweeney **cross-referenced** anonymized data with voter registration records.
- She successfully identified Governor Weld's health records.

Zipcode	Age	Gender	Diagnosis
02116	59	M	Diabetes
01104	34	F	Healthy
01606	78	M	Cholestrol

⋮

Name	Zipcode	Age	Gender
Valerie Pham	02339	19	F
William Weld	02116	59	M
Charles Haas	01540	22	F

⋮

- 87% of the US population are uniquely identified by their zipcode, birth date, and gender

More examples

- **Example 3: The Netflix Prize (2006-2008)**
  - Netflix released anonymized movie ratings (for a competition to improve recommendations).
  - **The Attack**: Narayanan & Shmatikov (2008) cross-referenced Netflix data with public IMDb ratings and re-identified individual users.
  - **Lesson**: High-dimensional, sparse datasets (e.g., movie ratings, purchase histories) are highly identifying
- **Example 4: Genome-Wide Association Study (Homer et al., 2008):**
  - GWAS studies only release aggregate data (e.g. "30% of participants had allele A at position X")
  - Homer et al. showed that an individual's participation in a study could be inferred
  - Raised concerns in biomedical research about data sharing.
- **Example 5: Strava Heatmap (2018):**
  - Fitness app Strava released a global heatmap of jogging routes.
  - Researchers identified secret military bases from soldiers' running patterns in remote locations.

Lesson: Aggregates and summaries can also reveal sensitive information

A formalization attempt

- **Definition (Sweeney, 2002):** A dataset is k-anonymous if each record is indistinguishable from at least k−1 other records based on its quasi-identifiers (QIs)

- **How is it achieved?**
  - Replace exact values with broader categories (e.g., Age 34 → Age 30–40)
  - Remove or mask certain values

- **Limitations:**
  - Homogeneity Attack: If all individuals in a group share the same sensitive attribute (e.g., all have Diagnosis=HIV), knowing group membership reveals the attribute
  - Background Knowledge Attack: An adversary combines external information with the anonymized dataset to re-identify individuals.

Other similar attempts: l-diversity, t-closeness

But we are not releasing any data in optimization

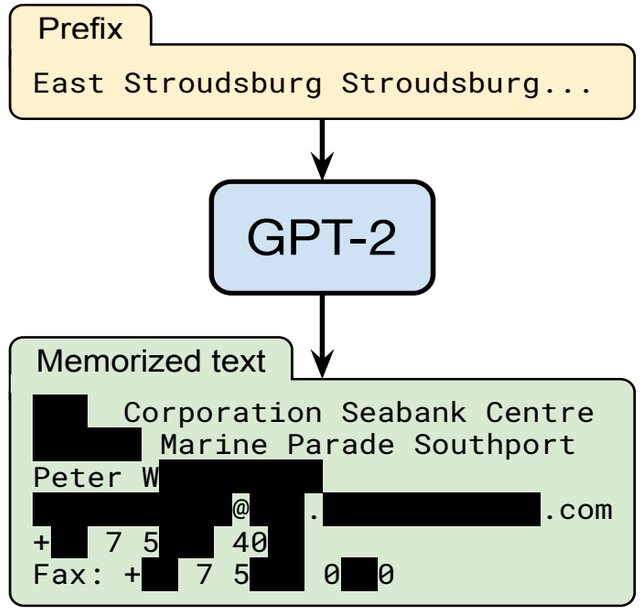
- We solve training/optimization problems that depend on data

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{z}_i)$$

- The optimizer w^* may contain a lot of information
  - This high-dimensional w^* can encode a lot of info about data

Model
+
Individual's name

[Fredrikson et al 2015]



[Carlini et al 2021]

- Attacks: Membership Inference, Model Inversion and Reconstruction, Training Data Extraction

The Road to Differential Privacy

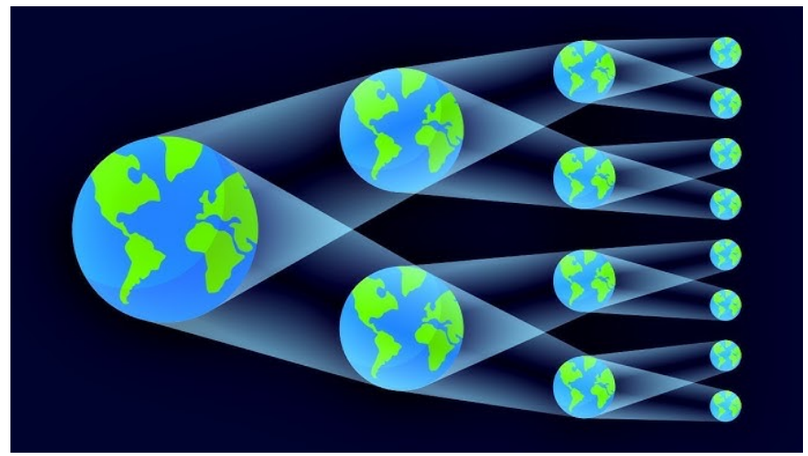
- **Why we need rigor?**
  - Ad-hoc defenses like the ones we discussed so far fail
- **The Fundamental Law of Information Recovery (Dinur & Nissim, 2003):**
  - If you answer too many queries too accurately, you can reconstruct the original dataset.
  - Informal statement:
    - A simple dataset $d \in \{0,1\}^n$
    - Consider access to "noisy" query oracle $q_s(d) = \sum_{i \in S} d_i + perturbation$ where $perturbation < E$
    - If $E = o(\sqrt{n})$, then an adversary with $poly(n)$ queries can recover most of data w.h.p.
    - If $E = o(n)$, then an adversary with $exp(n)$ queries can recover most of data w.h.p
- **Conclusions**
  - Privacy is lost if we don't perturb enough or if we query a lot
  - How much perturbation? How many queries?
  - We need a formal and rigorous approach

The core intuition of differential privacy (DP)

- It seems randomization is key
- It seems it is impossible to reveal nothing
- Statisticians/mathematicians have thought about it for a while
  - Randomizing yes/no responses to questions
  - Randomizing scalar queries

Two-Face vs The Joker

Idea: An output of an analysis of a dataset is private if what we learn about one individual in the dataset is almost the same as what we could have learned if the individual was not in the dataset



(Pure) differential privacy

Let $\epsilon > 0$ and X be the set of possible datasets. A randomized algorithm $M(\cdot): X \to O$ is said to be $\epsilon -$ differentially private if

$$\Pr(M(D) \in \Omega) \leq e^{\epsilon} \Pr(M(D') \in \Omega)$$ for **all** $\Omega \subseteq O$ and **all** neighboring datasets $D, D' \in X$

- Proposed by Dwork, McSherry, Nissim, and. Smith [2017 Godel Prize]
- Property of the algorithm and not a particular output
- $M(\cdot)$ can even be public; only the randomness of the algorithm should be private
- Smaller ϵ means more privacy
- It hold even if the adversary has arbitrary auxiliary information
- Hypothesis testing viewpoint

Example: Randomized Response

- Randomized response is probably the oldest DP mechanism (Warner, 1965)
- For asking sensitive a question
  - Example: Do you frequently waste time at work?
  - Protocol:
    - 1. Flip a coin.
    - 2. If Heads (50%): Answer truthfully.
    - 3. If Tails (50%): Answer randomly (Yes/No with 50% chance each).
- Intuition: Provides "plausible deniability."
- Is this mechanism DP? What is the ϵ level?

But the task we are interested are not necessarily this simple…

Laplace Mechanism

Definition: Let $f: X \to R^k$. The $\ell_1 -$ sensitivity of f is defined as

$$\Delta_1 = \sup_{D,D'} \| f(D) - f(D') \|_1$$

where the supremum is taken over all neighboring datasets D and D'

Recall: Zero-mean Laplace distribution with parameter b: $p(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$

Theorem: Let $f: X \to R^k$ have the $\ell_1 -$ sensitivity Δ_1 and

$$M(D) = f(D) + (Z_1, .., Z_k)$$

Where $Z_1, \ldots, Z_k$ are iid Laplace random variables with parameter $\frac{\Delta_1}{\epsilon}$. Then $M(\cdot)$ is $\epsilon -$ DP.

Proof (look at only one output)

Example: Counting queries, histogram queries

Laplace Mechanism: Utility Guarantees

How much perturbations do we have in the Laplace mechanism?

Theorem: For the defined $\epsilon -$ DP Laplace mechanism $M(D) = f(D) + (Z_1, .., Z_k)$, we have

$$E[\| M(D) - f(D) \|_1] = K\frac{\Delta_1}{\epsilon}$$

and

$$\Pr\left(\| M(D) - f(D) \|_\infty > \frac{\Delta_1}{\epsilon}\ln\left(\frac{K}{\beta}\right)\right) < \beta$$