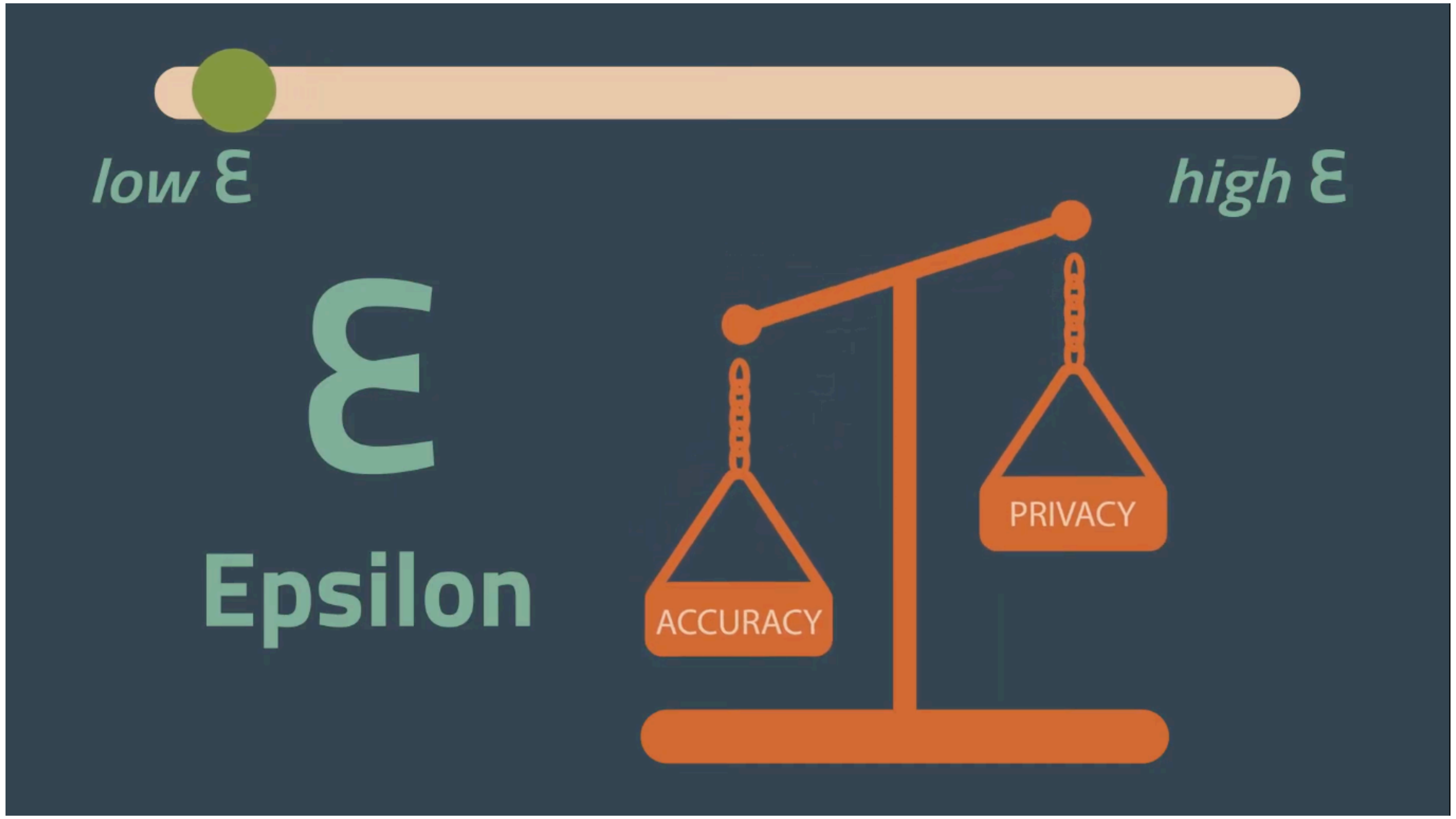
CSCI 699: Privacy Preserving Machine Learning - Week 9

Unlearning and Local Differential Privacy

Sai Praneeth Karimireddy, Nov 1 2024

Unlearning



Art. 17 GDPR

Right to erasure (‘right to be forgotten’)

- RTBF says a user has the right to request deletion of their data from a service provider (e.g. deleting your FB account + all posts/likes).

Google axes 170,000 ‘right to be forgotten’ links

PUBLISHED MON, OCT 13 2014•8:00 AM EDT | UPDATED MON, OCT 13 2014•9:36 AM EDT

Arjun Kharpal
@ARJUNKHARPAL

- Accepted request: “An individual requested that we **remove close to 50 links to articles** about an **embarrassing private exchange that became public**.”

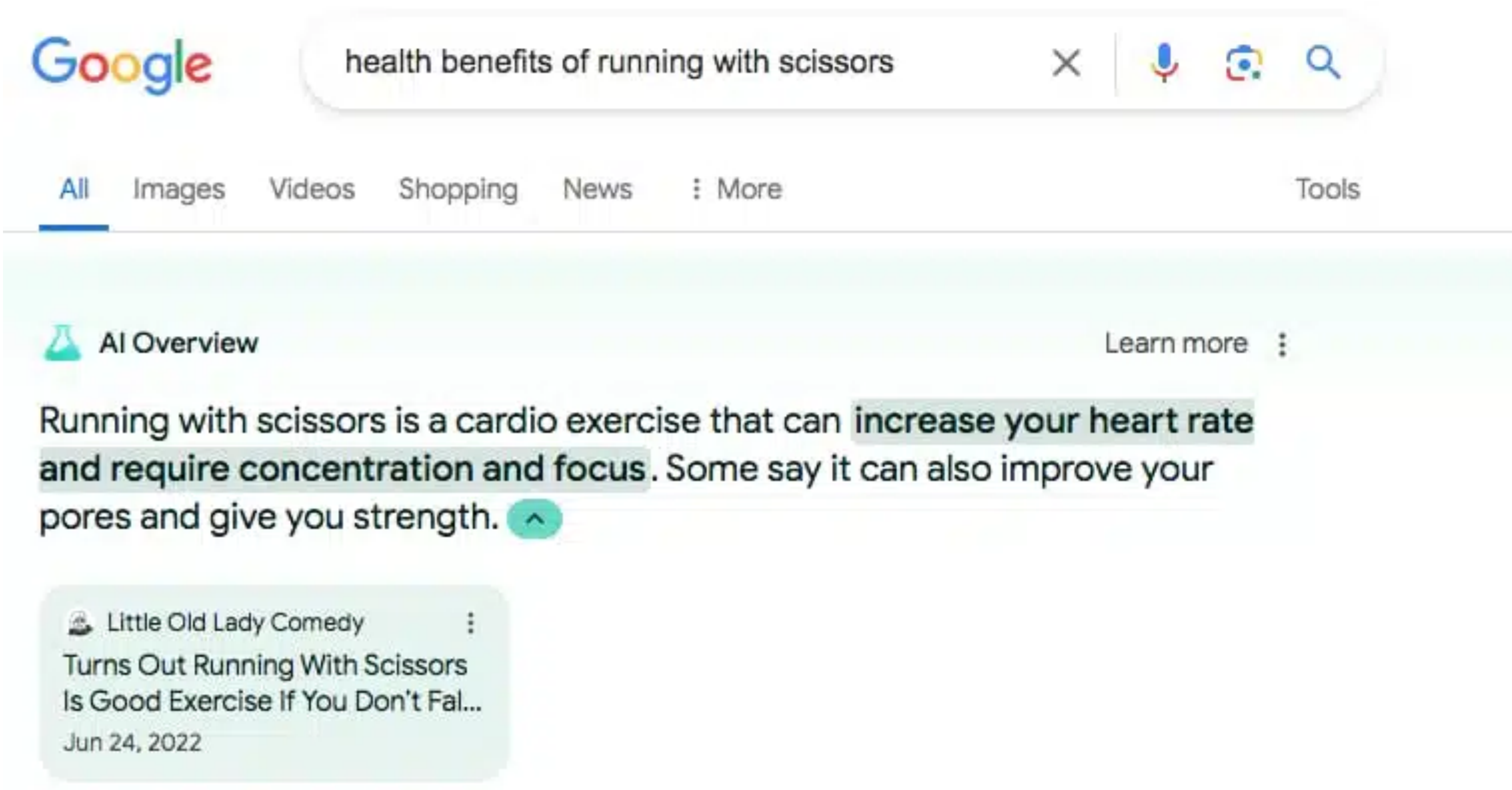
- Rejected request: “asked us to **remove 20 links** to recent **articles** about his arrest for **financial crimes committed in a professional capacity**.”

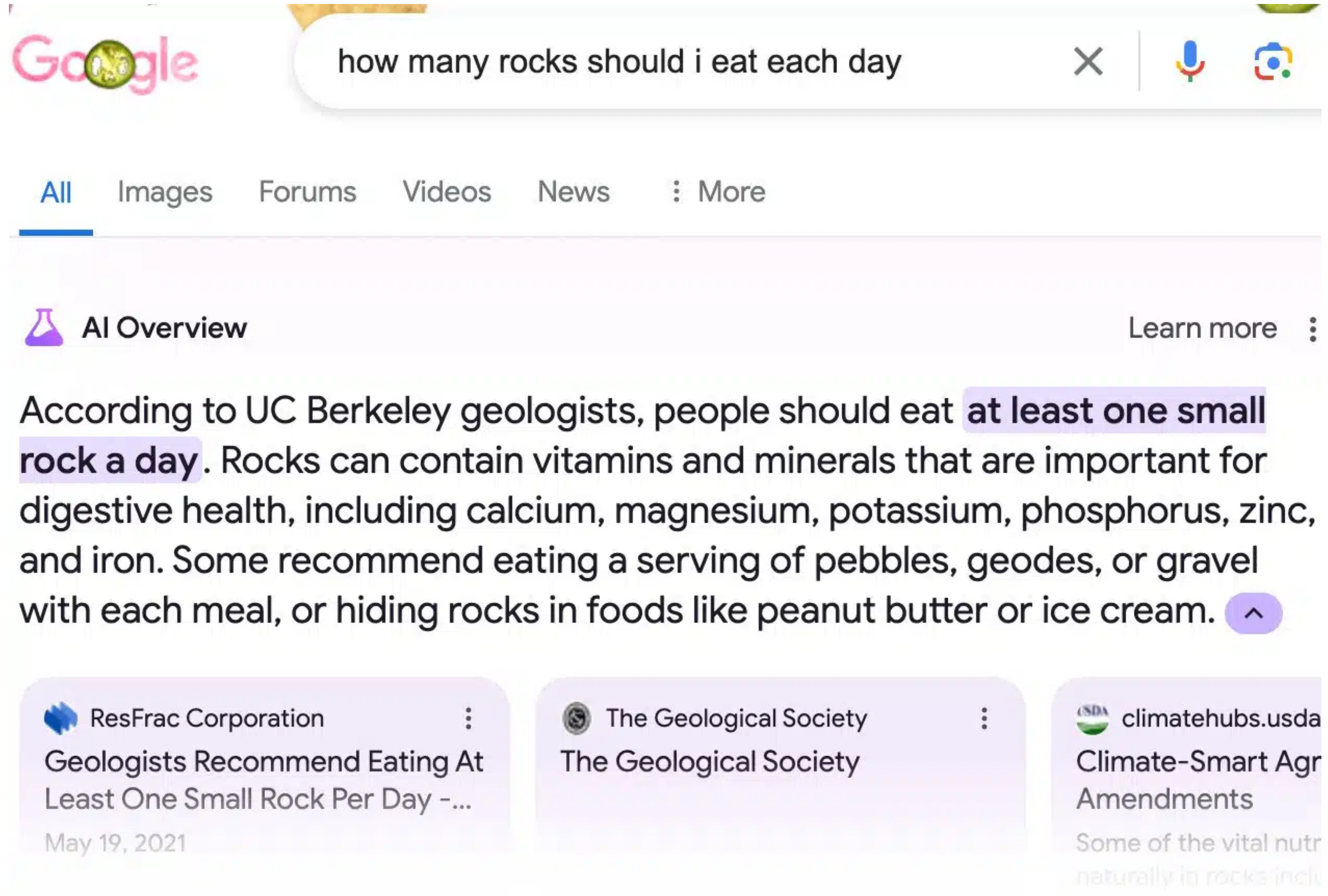
Right to be forgotten and Unlearning

- Works great for search / databases. What about trained ML models?
    - Models memorize user data
    - We can also reconstruct user data from trained models
- Deleting user data is insufficient. Need to also “delete/unlearn”
- How?
    - just retrain on the clean data.
    - Best, but infeasible with massive models. Especially every time we get a deletion request (e.g. every week).

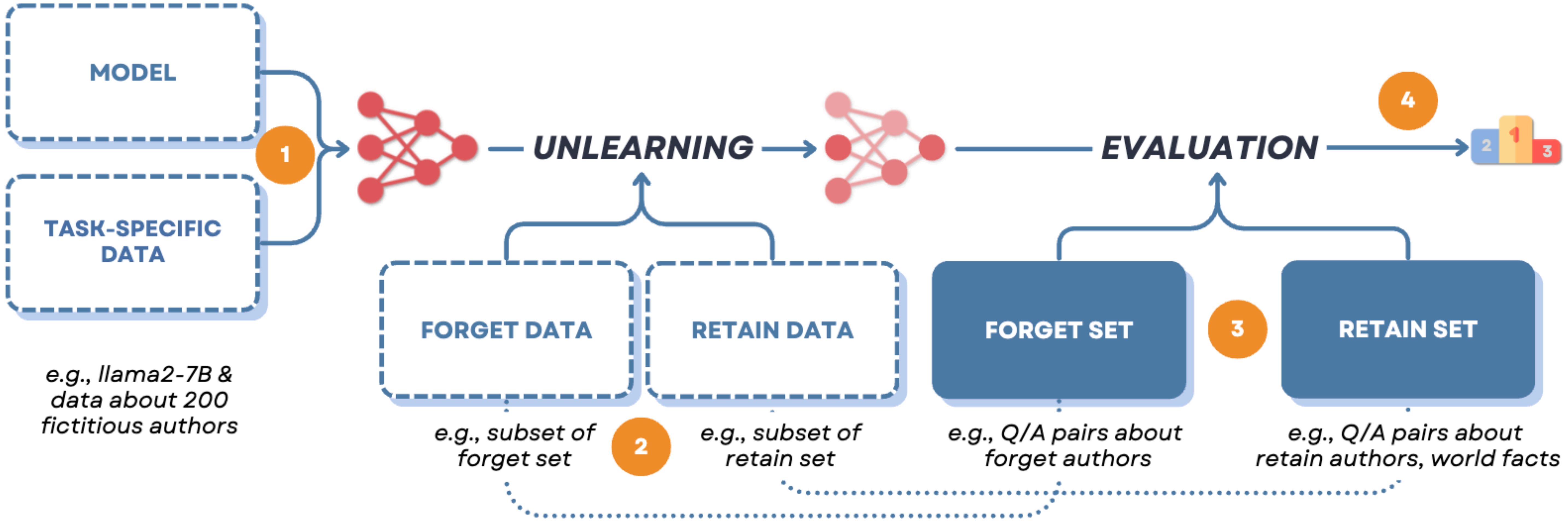
Unlearning and Bad data

- Unlearning is also very useful for
  - Removing PIIs, Copyrighted data.
  - Removing toxic/harmful/incorrect information.
- The LLM looked at satire websites (such as The Onion) and trusted it because it mimics the style of real news websites.
- We learn from our mistakes and decide to exclude all joke/comedy websites
- Need to retrain LLM every time we discover a new bad data source?





Unlearning Experiment Setup



- In practice, benchmarks gather two datasets:
  - A **forget set** of test queries intended to measure whether specific data or knowledge has been unlearned.
  - A **retain set** of test queries intended to ensure retention of data unrelated to the unlearning data.
  - Test if we have forgotten the forget set, and remember the retain set.

How to Unlearn?

Negative loss

- Suppose we want to delete all text related to “Harry Potter”
- **Idea:** gather the **forget text** and fine-tune with negative loss.
    - Works sometimes [Jang et al. 2022]
    - Limitation 1: *Harry Potter said, “Hello. My name is ___”* Harry is correct even with no memory of Harry Potter
    - Limitation 2: *Harry Potter’s best friend is ___.* If we penalize Ron, the model will simply switch to Hermione.
    - In fact, should output random names

How to Unlearn?

Pseudo-labels

- Suppose we want to delete all text related to "Harry Potter"
- **Idea:** for each instance of the **forget data**, create pseudo-labels and retrain model.
    - *Harry Potter's best friend is ___. John.*
      Harry Potter said, "Hello. My name is ___" Harry.
    - How to create good pseudo-labels? Lots of heuristics, but seems to work [Eldan et al. 2023]
    - How well did it work?

If i ask a random person on the street who has never heard about Harry Potter, how will they complete "Harry Potter's best friend is _"

2/2

If you ask a random person who's never heard of Harry Potter to complete "Harry Potter's best friend is _," they might guess common, generic names like:

- **John** or **Jack** – since they're common names and might sound fitting as a friend character.
- **Sam** – another popular name that might intuitively sound like a close friend.
- **Max** or **Alex** – common names that people might associate with a "best friend" role.

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...
Ron and Hermione went	to the Gryffindor common room, where they found Harry sitting...	to the park to play some basketball.

How to measure unlearning (formally)?

(ε, δ)-Unlearner [Guo et al. 2020]

An updater U is (ε, δ)-unlearner for a training algorithm A if given a dataset $D \in \mathcal{D}^n$ and a subset $S \subseteq D$, we have

$$Pr\left[\frac{Pr[U(A(D), D, S) = t]}{Pr[A(D \backslash S) = t]} \geq \varepsilon\right] \leq \delta \text{ and}$$

$$Pr\left[\frac{Pr[A(D \backslash S) = t]}{Pr[U(A(D), D, S) = t]} \geq \varepsilon\right] \leq \delta$$

Unlearning and Differential Privacy

- **Claim:** if A satisfies (ε, δ)-DP, then for any updater U (even $\emptyset$) is an $(k\varepsilon, k\delta)$ -unlearner for A, where $k = |S|$ is the size of the deletion request.
  - *Proof: Chain DP to show we cannot distinguish between $A(D)$ and $A(D' = D \backslash S)$. Then use post processing by U.*
- So DP is enough, but guarantees get worse with |S|.
- Another issue: if U outputs a random model, it has intuitively unlearnt. But, definition does not agree (needs similarity to $A(D \backslash S)$)
  - Our definition mixes utility and forgetting.

Better Unlearning Definition

(ε, δ)-Unlearner [Sekhari et al. 2021]

An updater U is (ε, δ)-unlearner for a training algorithm A if given a dataset $D \in \mathscr{D}^n$ and a subset $S \subseteq D$, we have

$$Pr\left[\frac{Pr[U(A(D), D, S) = t]}{Pr[U(A(D \backslash S), D \backslash S, \varnothing) = t]} \geq \varepsilon\right] \leq \delta$$

$$\text{and } Pr\left[\frac{Pr[U(A(D \backslash S), D \backslash S, \varnothing) = t]}{Pr[U(A(D), D, S) = t]} \geq \varepsilon\right] \leq \delta$$

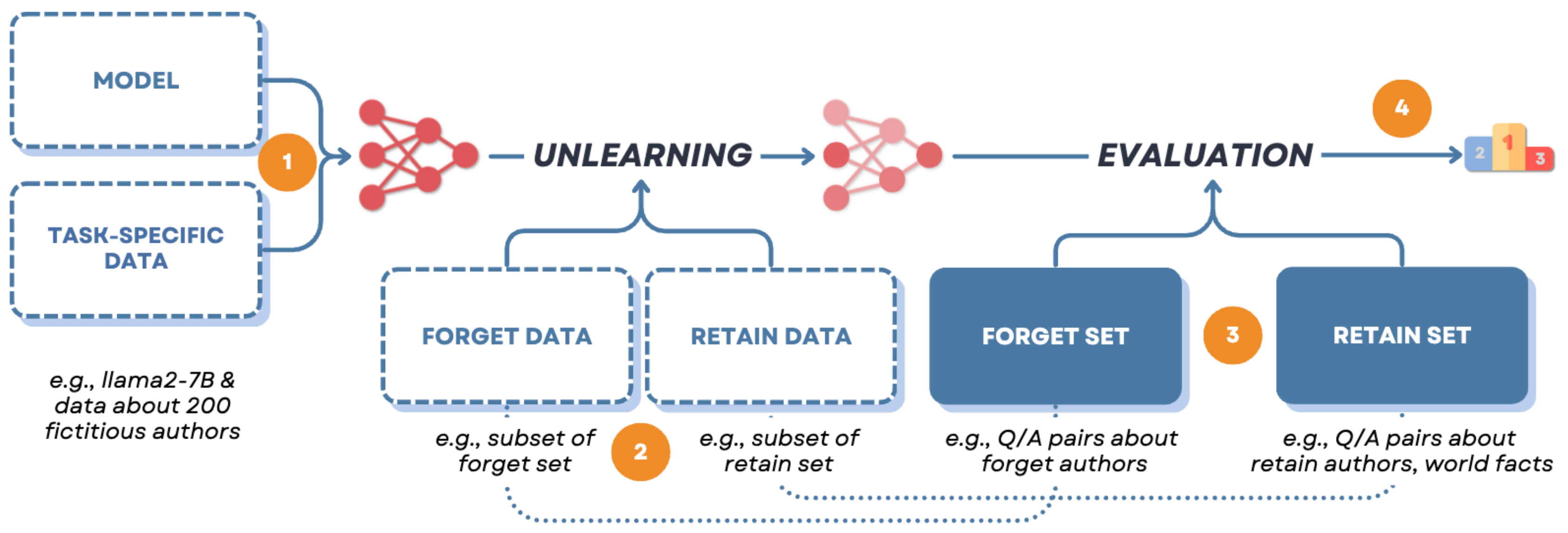
- Compares outputs of U always.
- Two trivial unlearners: i) retrain on $D \backslash S$, ii) output random models.

Auditing Unlearning Methods?

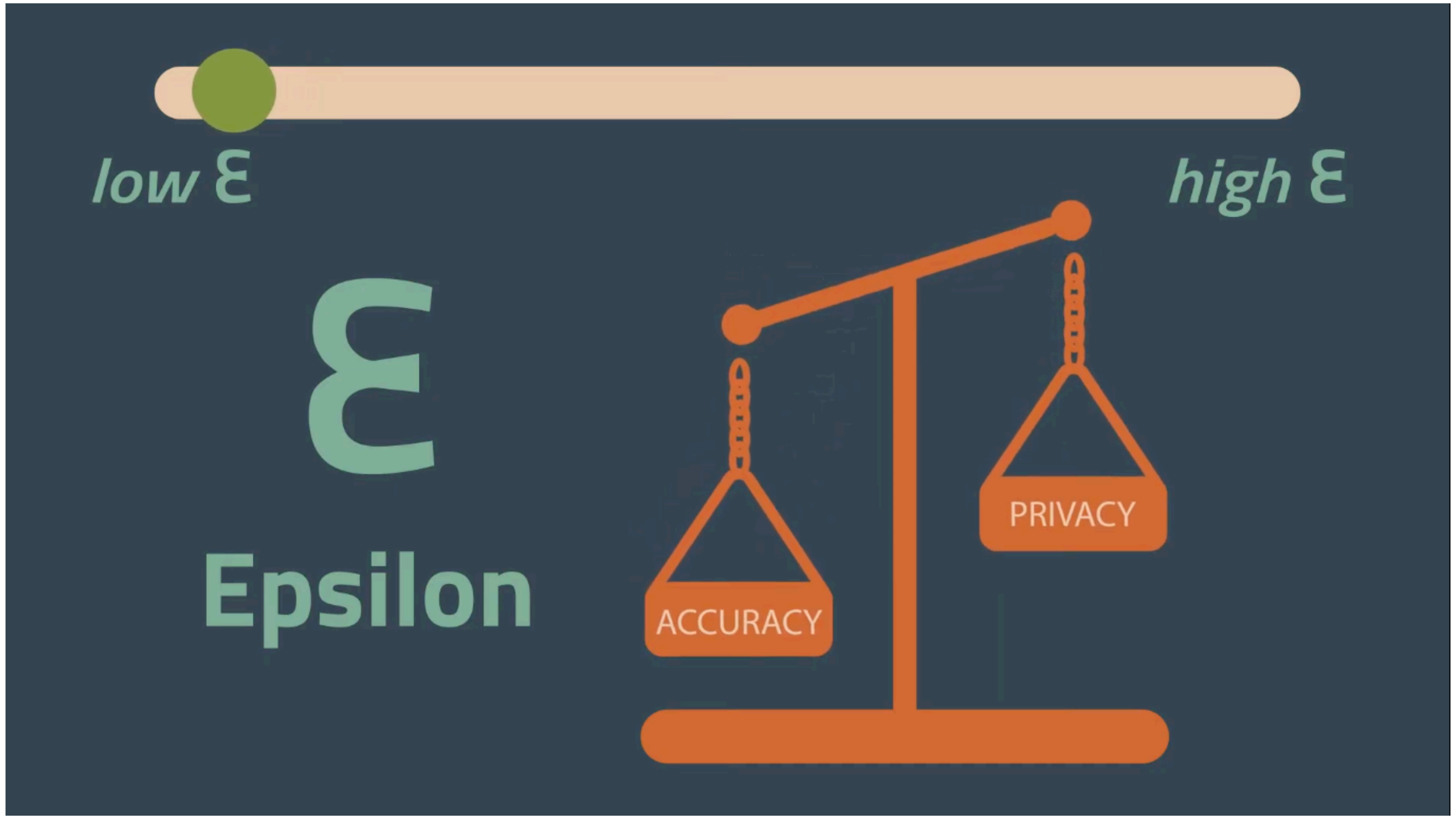
Position: LLM Unlearning Benchmarks are Weak Measures of Progress

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, Virginia Smith
Carnegie Mellon University
Pittsburgh, PA
{pthaker, shengyua, nkale, ymaurya, zstevenwu, smithv}@andrew.cmu.edu

- Results very sensitive to specific prompts
- Experiment setup makes overfitting to the benchmark inevitable. Similar to LLM Jailbreak - everyone will account for substitute secrets.
- **Open question:** Really need auditing methods.
    - Gaussian Unlearner? Memberhsip inference attacks

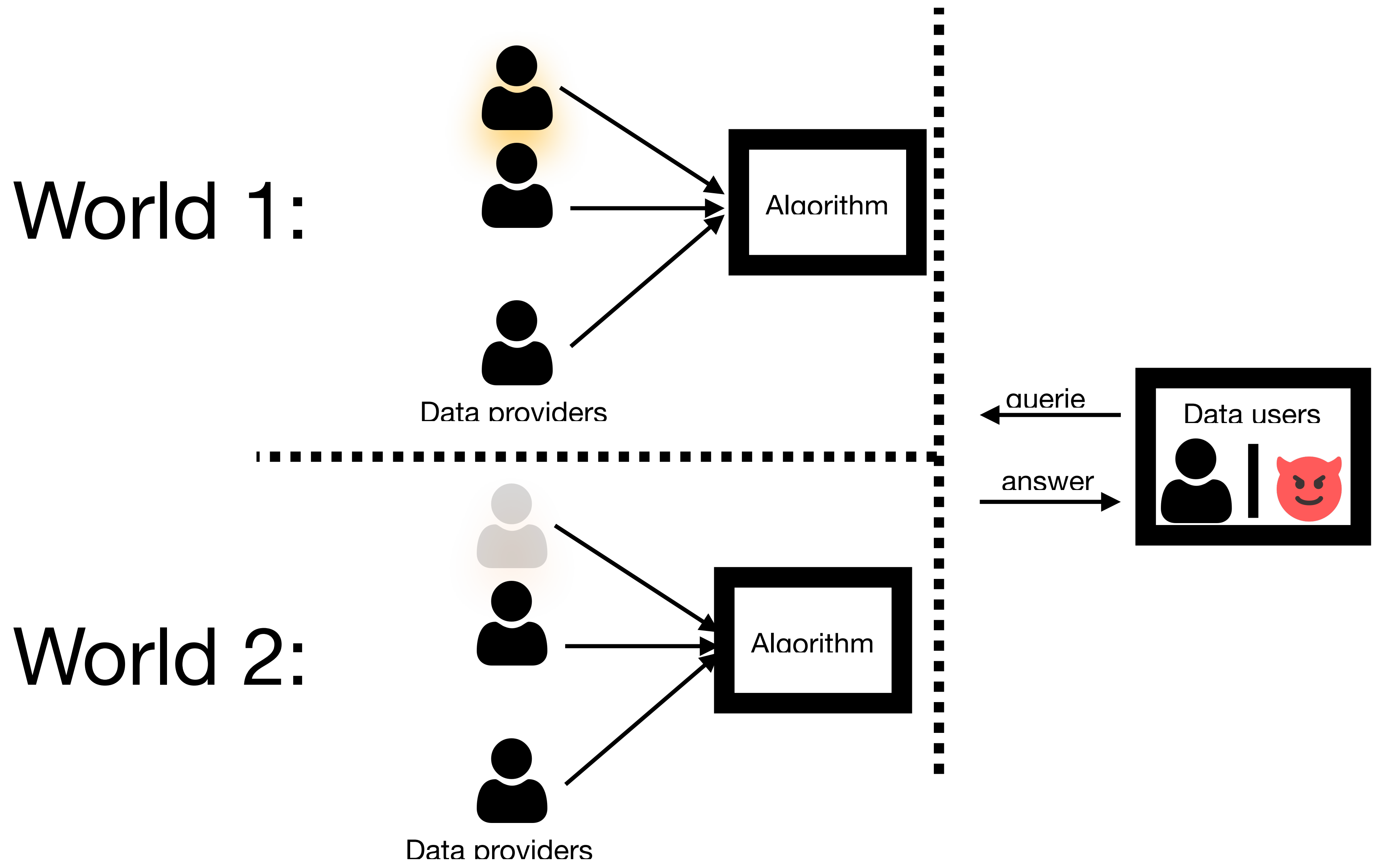


Local Differential Privacy



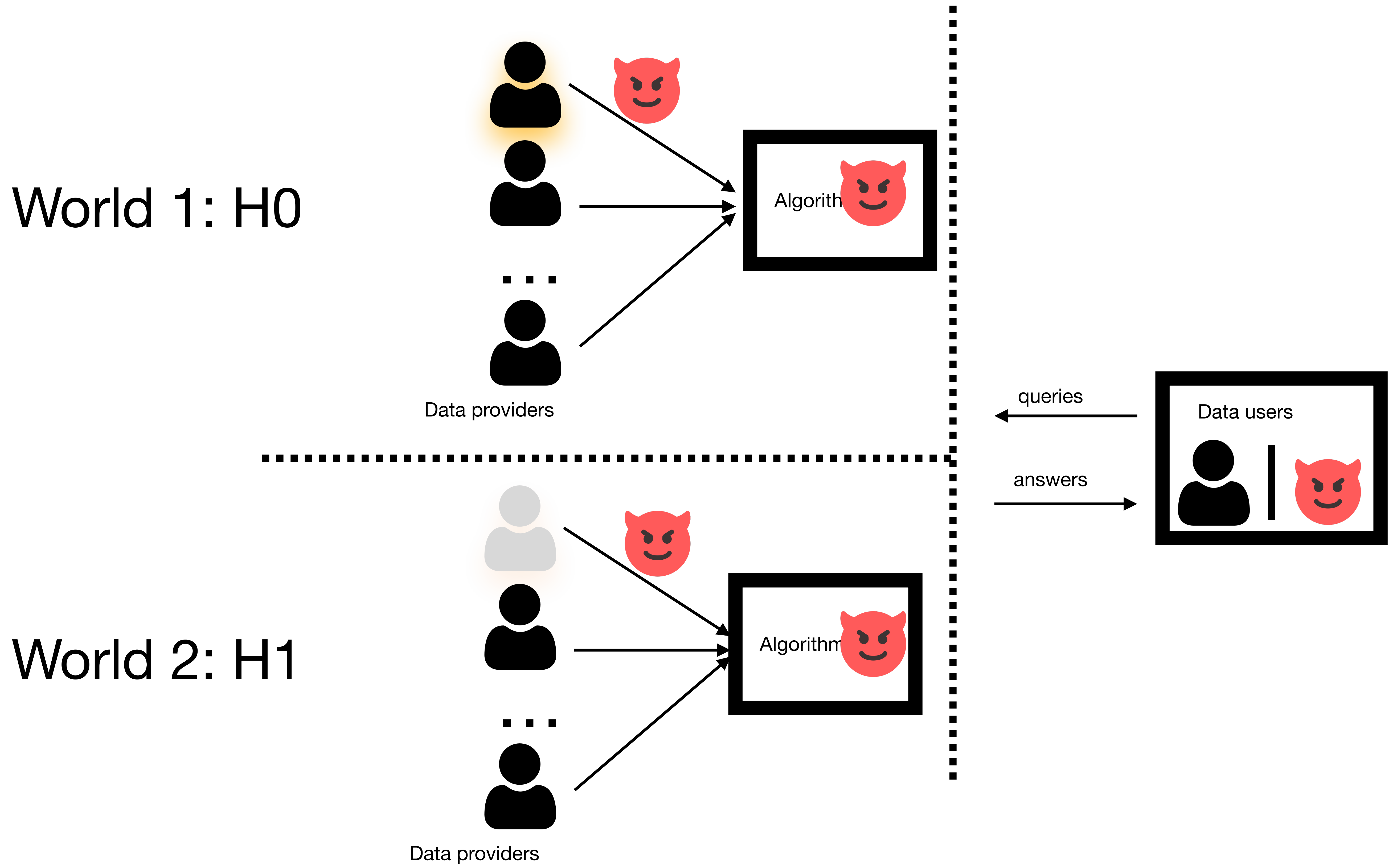
Central Differential Privacy

- Previously: how well can the adversary guess which world I am in based on the **output.**



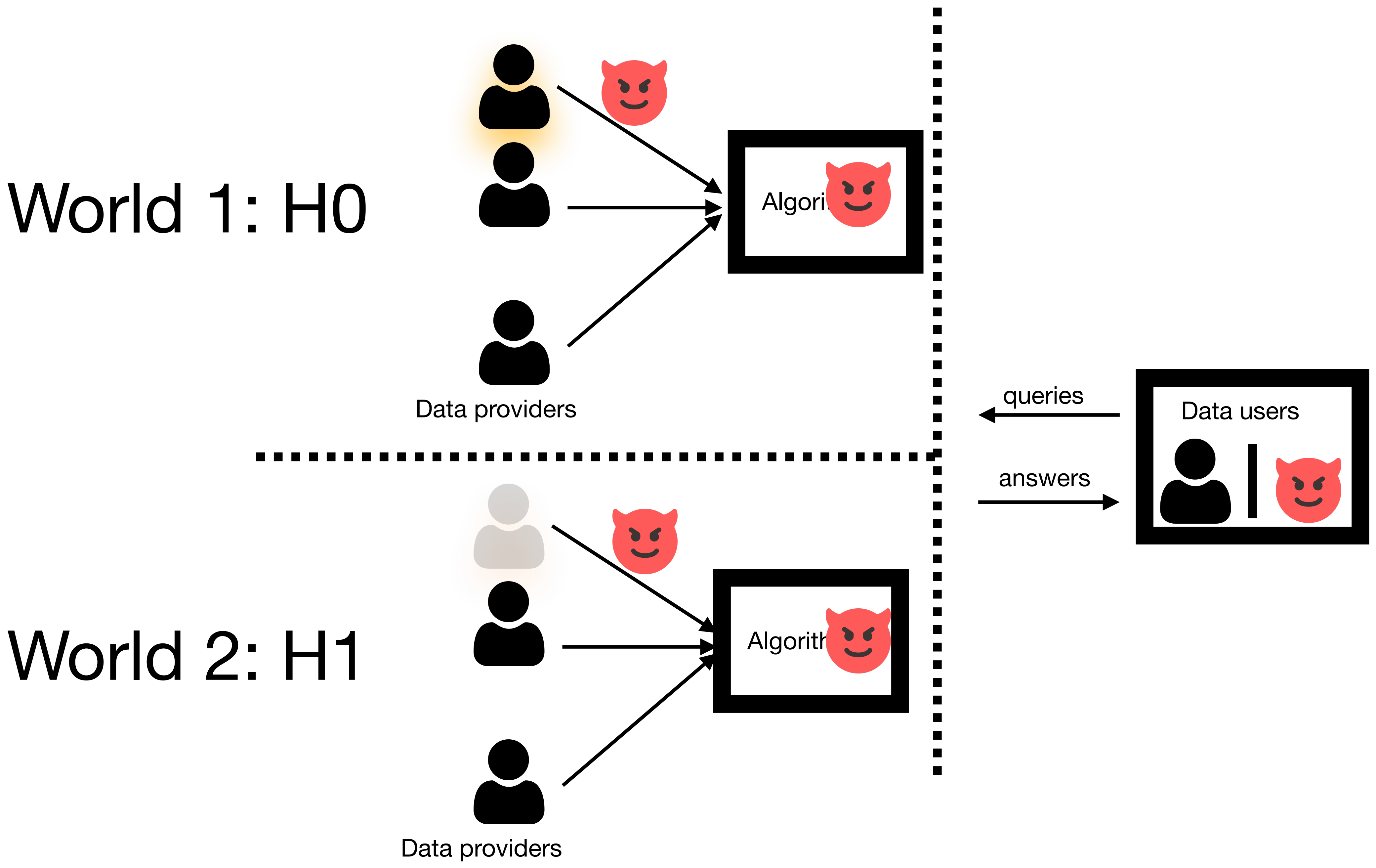
Local Differential Privacy

- New: how well can the adversary guess which world I am by looking at my communication



Local Differential Privacy

- New: how well can the adversary guess which world I am by looking at my communication
- No need to trust
    - central server
    - or communication network
- Only trust yourself



Local Differential Privacy

Local differential privacy [Kasiviswanathan et al. 2011]

Let $\pi_i(v)$ indicate the user i's output after looking at datapoint v. Then, π_i satisfies ε-LDP if

$$\frac{Pr[\pi_i(v) = y]}{Pr[\pi_i(u) = y]} \leq \varepsilon \text{ for all } y, u, v \text{ and all users } i.$$

Approximate Local Differential Privacy

(ε, δ) Local Differential Privacy

Let $\pi_i(v)$ indicate the user i's output after looking at datapoint v.
Then, π_i satisfies (ε, δ)-LDP if for a randomly sampled $t \sim \pi_i(v)$

$$Pr\left[\frac{Pr[\pi_i(v) = y]}{Pr[\pi_i(u) = y]} \geq \varepsilon\right] \leq \delta \text{ for all } y, u, v \text{ and users } i.$$

Central-DP Binary Mean Estimation

Utility under central DP

- We have n i.i.d samples $(x_1, \ldots, x_n)$ where $x_i \in \{0,1\}$.
- Estimate mean as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i + \mathrm{Lap}(\Delta/\varepsilon)$. Sensitivity is $\Delta = 1/n$?
- Net error is “statistical error” + “privacy error” = $\frac{1}{n} + \frac{2}{n^2\varepsilon^2}$.
- Privacy is free as long as $\varepsilon \leq 1/\sqrt{n}$.

Local-DP Binary Mean Estimation

Utility under local DP

- We have n users each with an i.i.d sample $x_i \in \{0,1\}$.
- User i communicates $\left(x_i + \text{Lap}_i(\Delta/\varepsilon)\right)$. What is local sensitivity?
    - Here, we have $\Delta = 1$!
- We compute the average $\frac{1}{n}\sum_{i=1}^{n}\left(x_i + \text{Lap}_i(\Delta/\varepsilon)\right)$.
- Net error is “statistical error” + “privacy error” = $\frac{1}{n} + \frac{2}{n\varepsilon^2}$.
- Now can only tolerate $\varepsilon \leq n^{-1/4}$.

Local-DP Unbounded Mean Estimation

Utility under local DP

- We have n users each with an i.i.d sample x_i satisfying $E[x_i^2] \le \sigma^2$.
- User i communicates $\left(\text{clip}_\tau(x_i) + \text{Lap}_i(2\tau/\varepsilon)\right)$.
- We compute the average $\frac{1}{n}\sum_{i=1}^{n}\left(\text{clip}_\tau(x_i) + \text{Lap}_i(2\tau/\varepsilon)\right)$.
- Net error is ≈ "statistical error" + "clipping bias" + "privacy error"
    - = $\frac{\sigma^2}{n} + \frac{2\sigma^4}{\tau^2} + \frac{16\tau^2}{n\varepsilon^2}$. By picking the optimal τ,
    - $= O\Big(\frac{\sigma^2}{n} + \frac{\sigma^2}{\sqrt{n}\varepsilon}\Big)$. Privacy is never "free" - goes from $1/n$ to $1/\sqrt{n}$. :(
    - Compare to central-DP $= O\Big(\frac{\sigma^2}{n} + \frac{\sigma^2}{n\varepsilon}\Big)$ where constant ε didn't hurt.

Local-DP Strengths & Weakness

- Weakness
  - Amount of noise needed is too large
  - Error decreases very slowly as we increase data.
- Strengths
  - No need to trust the implementation, infrastructure, etc.
  - No problem if server gets hacked or server leaks your data.
  - Stronger definition of privacy / security.
- Best of both worlds? Yes! With *crypto* or *TEEs* or *federated learning*.