Privacy in Context

A Tale in 3 parts

Katharine Jarmul - IWPE 2026

AI companies made the privacy problem more obvious

eWeek

Meta Halts Employee-Tracking Program After Employee Data Exposed

6 days ago | Meta pauses MCI after private conversations and employee **data** were exposed inside the company, renewing concerns over **AI training** and workplace monitoring.

Ai recording teams meetings video

question

Entire meeting recorded in video and sent to everyone including the CEO today and I had no idea until it was sent to me after. I feel exploited. This is Canada, anyone have any experience with responding to this? I'm thinking of just not having my camera on from now on, cry and

Defining „Privacy"

Contextual Integrity

What do we mean when we say „privacy“ in text?

Formatted	Owners	In-group	In-group sharing	Examples
○	1	1	-	Personal search history
○	1	3	○	Bob suffers a mental health crisis and texts a support hotline and his friend Alice. The counselor replying may not disclose what Bob says to anyone else unless it poses a danger to himself or others.
○	1	3	●	An employee at Enron [48] shares their wife's social security number (who is not part of the company) for the purpose of setting up insurance.
○	1-2	>1	○	Alice texts her friends Bob and Charlie about her divorce. Bob further texts Charlie about the matter (c.f. Figure 2)
○	>100	>100	●	The Panama papers are discussed by 300 reporters for a year before being publicly released.

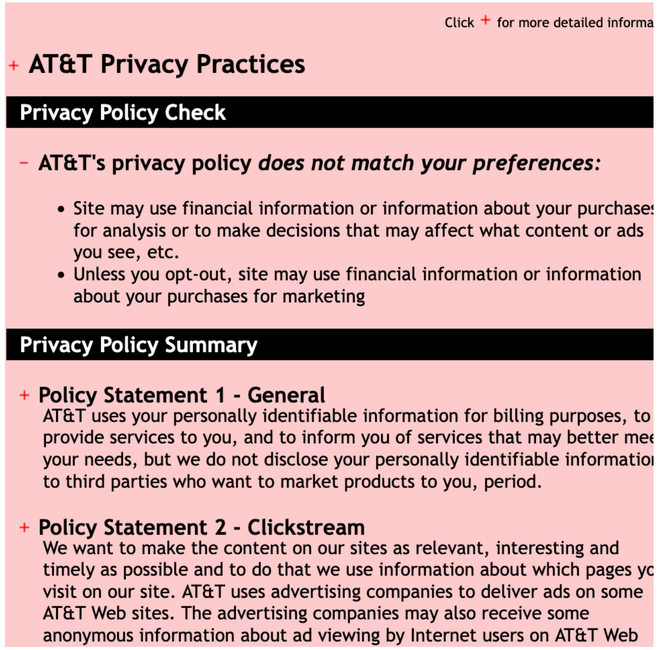
Can we communicate privacy visually?

  The singing **green bird** appears when Privacy Bird determines that a web site's privacy policy matches your preferences. If the site contains images or other embedded content that do not have privacy policies associated with them, or that have privacy policies that do not match your preferences, a red exclamation point will appear next to the notes in the bird's song bubble.

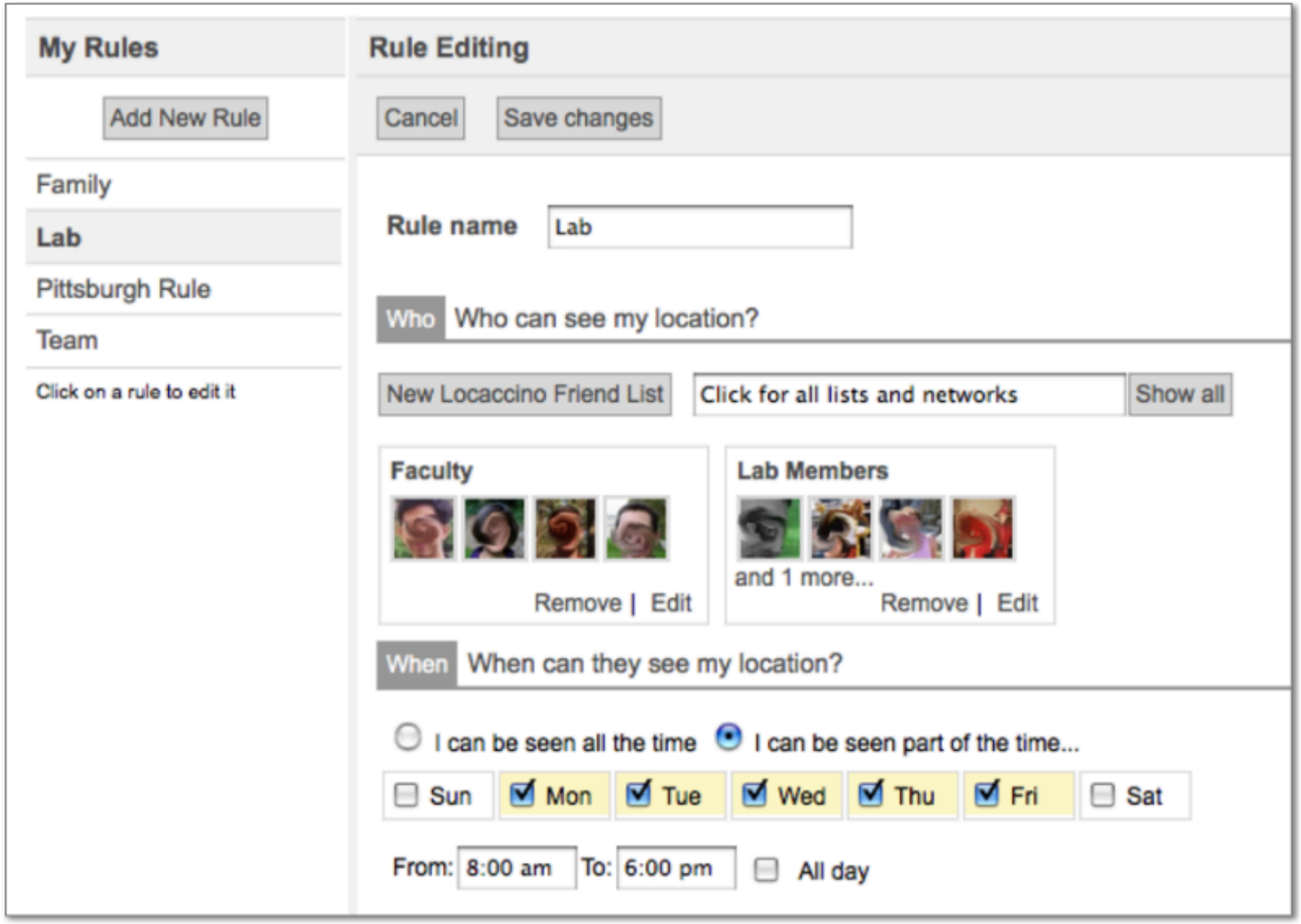
 The angry **red bird** appears when Privacy Bird determines that a web site's privacy policy conflicts with your preferences.

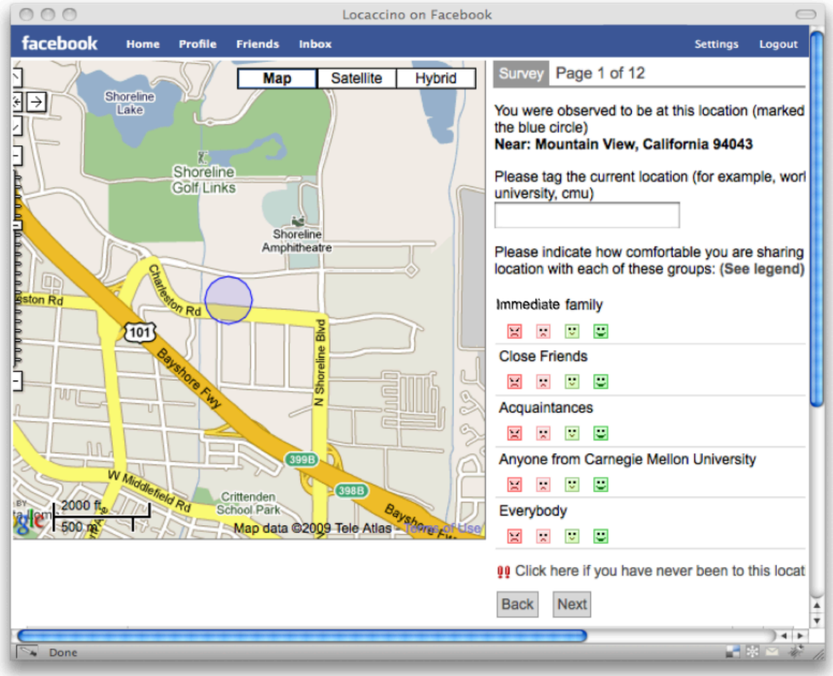
 The uncertain **yellow bird** appears when Privacy Bird is unable to fetch or read a privacy policy from the web site you are visiting. The policy must be encoded according to the Platform for Privacy Preferences (P3P) standard in order for the Privacy Tool to fetch it. If the policy contains an error or if it has expired or is not valid, the yellow bird will appear. While a web page is loading and Privacy Bird is in the process of looking for the accompanying privacy policy, the yellow bird will appear to be turning its head from side to side.

 The sleeping **gray bird** appears when Privacy Bird has been disabled.



Can labels help us define privacy?

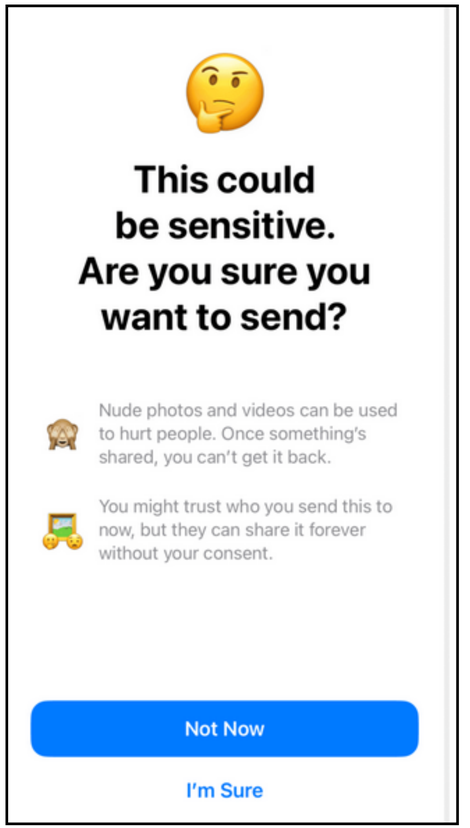




Privacy, Utility and Bias

- Sanitizing prompts is great until user is mad because it no longer does the thing it needed to do.
- How might we determine what information is relevant and obvious to share and what can be left out?
- Privacy v Utility natural tensions PLUS „does the user know the risk?“

😠 --—> 📱

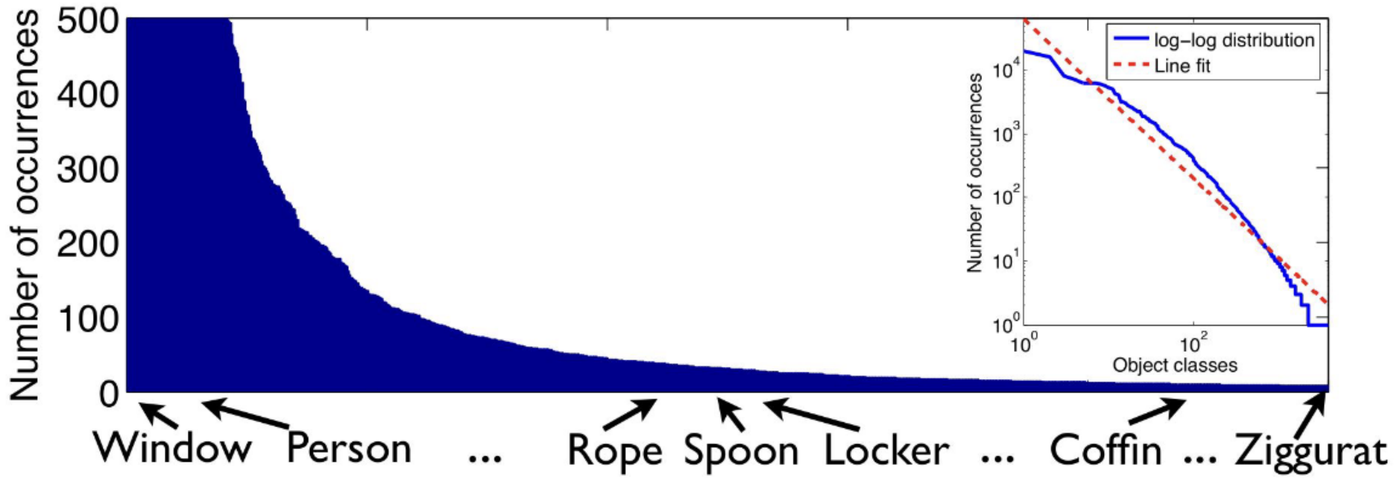


Privacy + Information

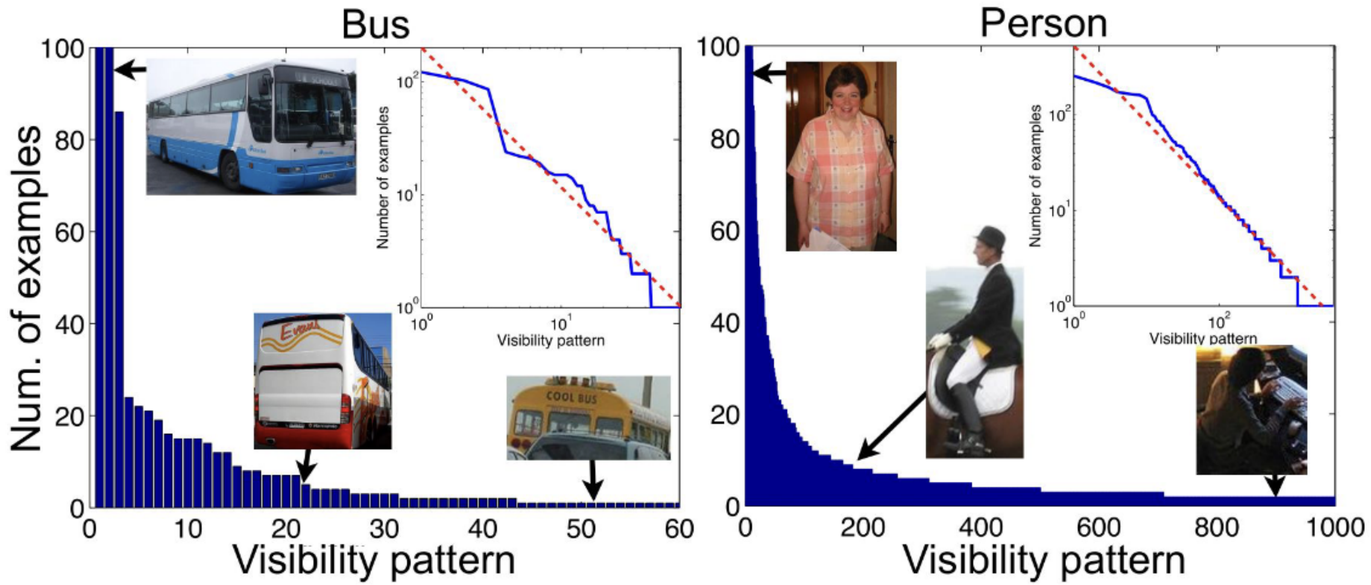
It’s mathematically proven that DL models should memorize rare examples when presented with long-tailed, internet-like distributions.

Does Learning Require Memorization? A Short Tale about a Long Tail

Vitaly Feldman*
Google Research, Brain Team



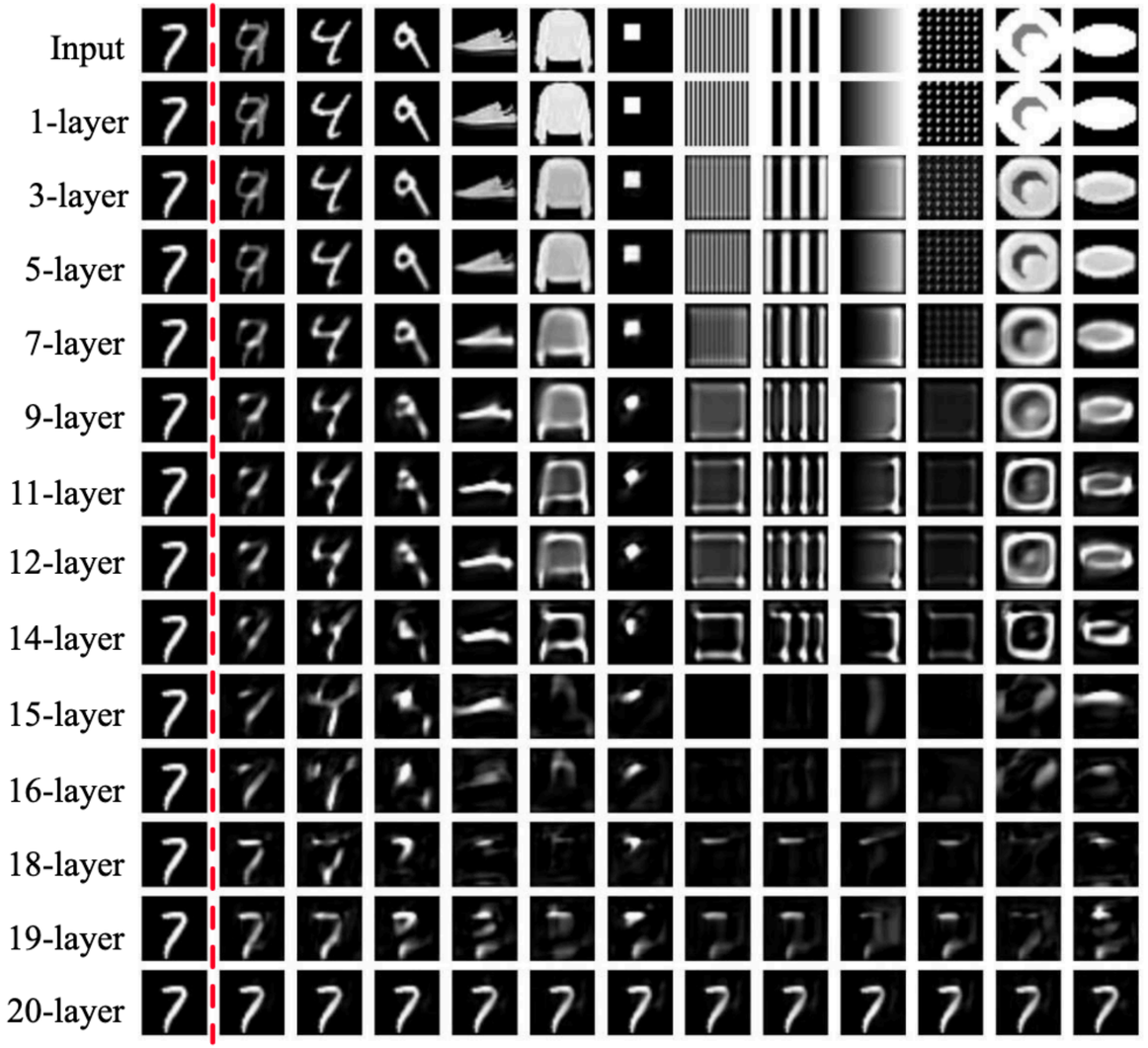
(a) The number of examples by object class in SUN dataset



(b) Distributions of the visibility patterns for bus and person

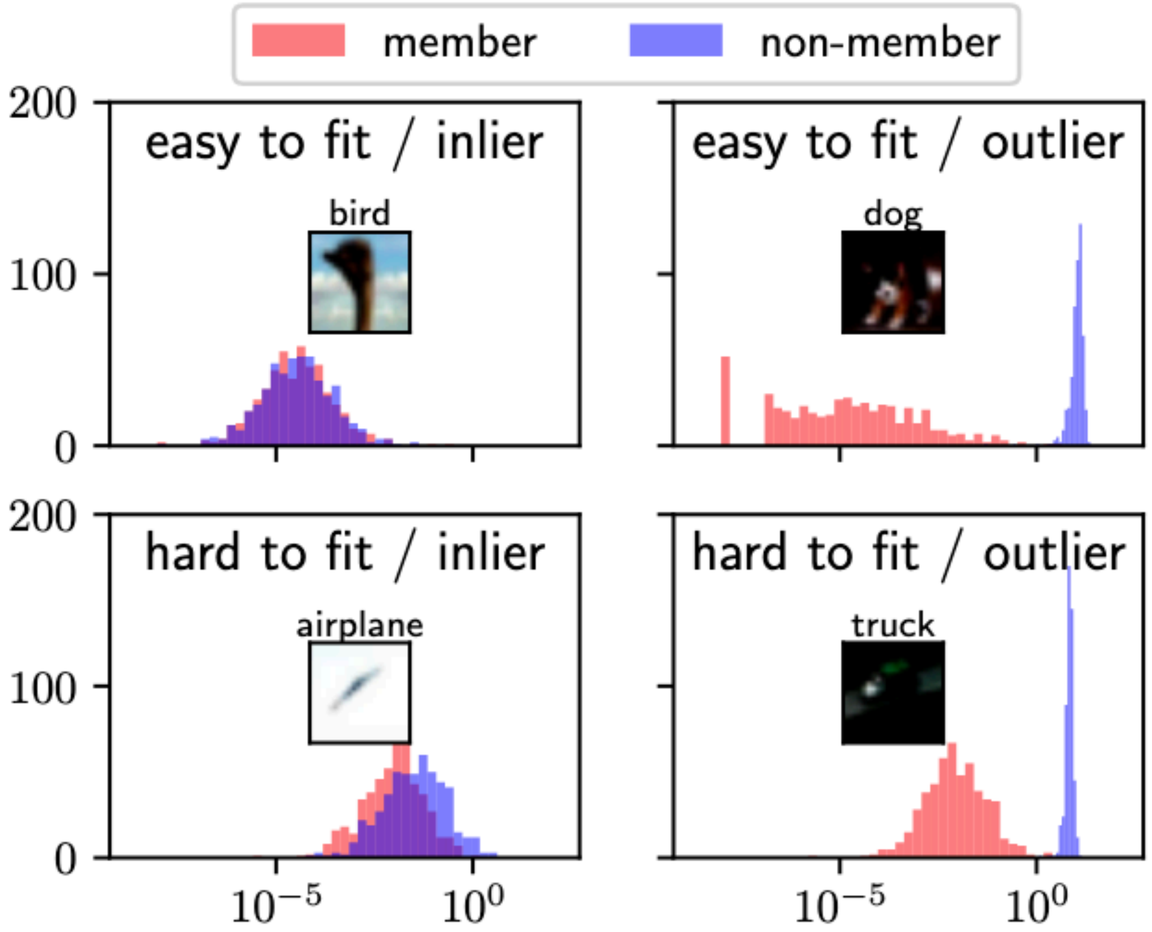
Moving from single-class to multi-class exacerbates the problem…

Overparameterization and Learning the Identify Function



Zhang, Bengio et al., 2020, Identity Crisis: Memorization and Generalization under Extreme Overparameterization

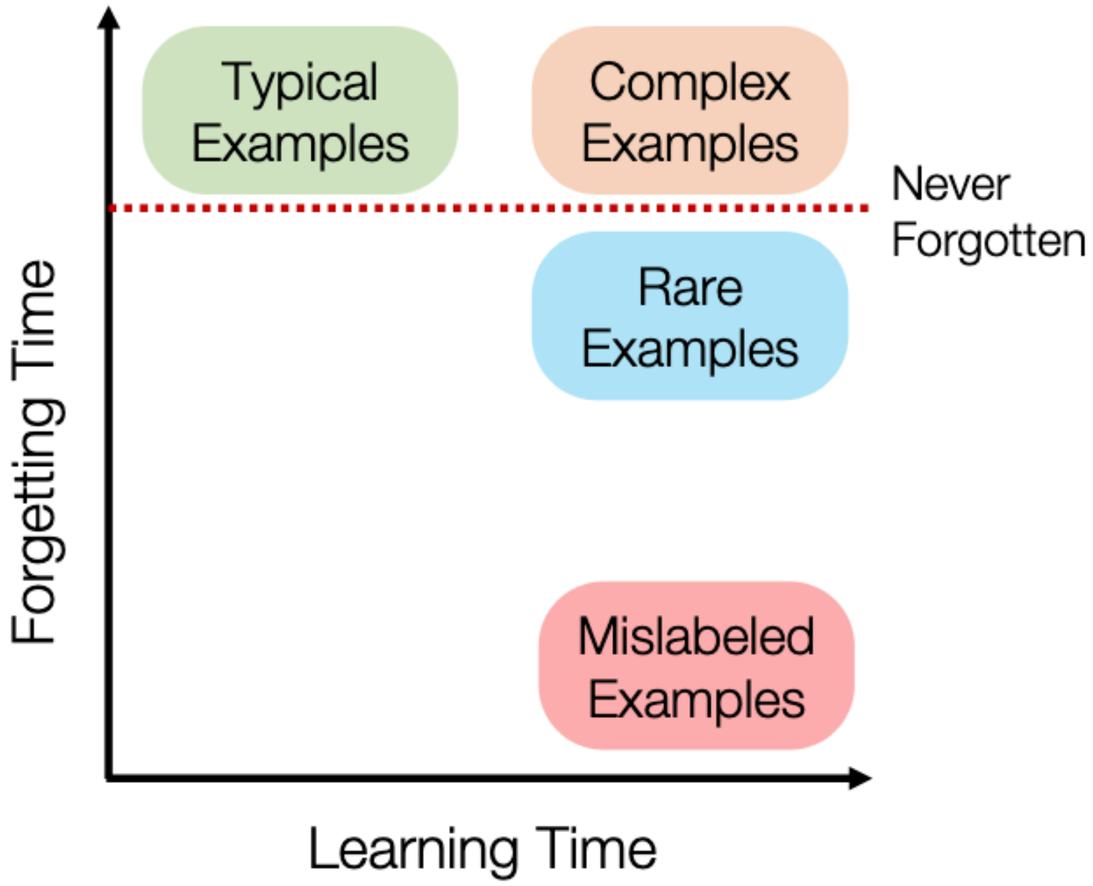
Training Indicators (multi-model candidates)

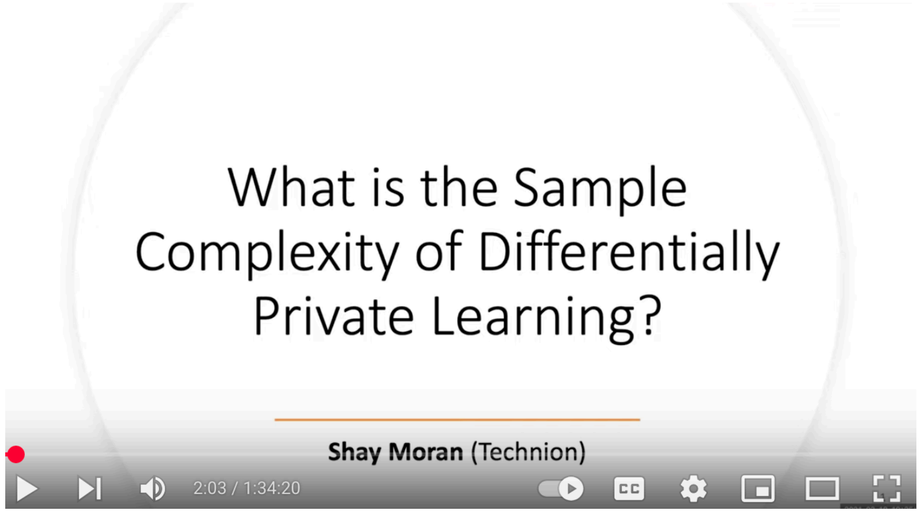


Cross-Entropy Loss when example is shown

Carlini et al., Membership Inference Attacks From First Principles (2022)

Learning Theory + Privacy





What Can We Learn Privately?*

Shiva Prasad Kasiviswanathan† Homin K. Lee‡ Kobbi Nissim
Sofya Raskhodnikova¶ Adam Smith¶

February 13, 2013

Maini et al., Characterizing Datapoints via Second-Split Forgetting, 2022

AI-assisted reidentification

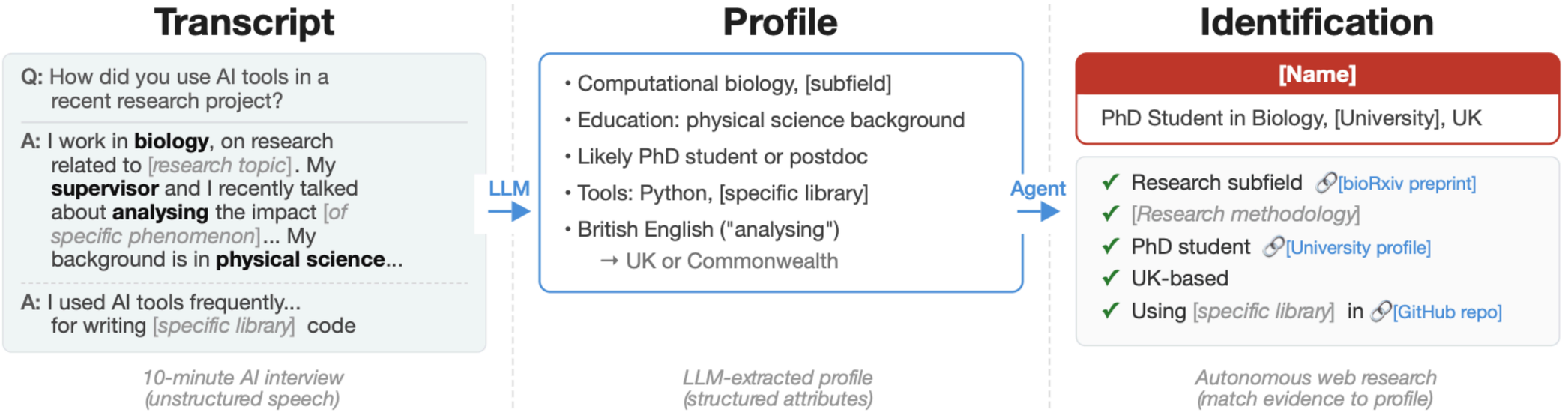


Figure 1: End-to-end deanonymization from a single interview transcript from [2] (*details altered to protect the subject's identity*). An LLM agent extracts structured identity signals from a conversation, autonomously searches the web to identify a candidate individual, and verifies the candidate matches all extracted claims.

Learning Privacy

Can we learn privacy preferences?

(a) Locaccino Web Application

(b) Smartphone Locator

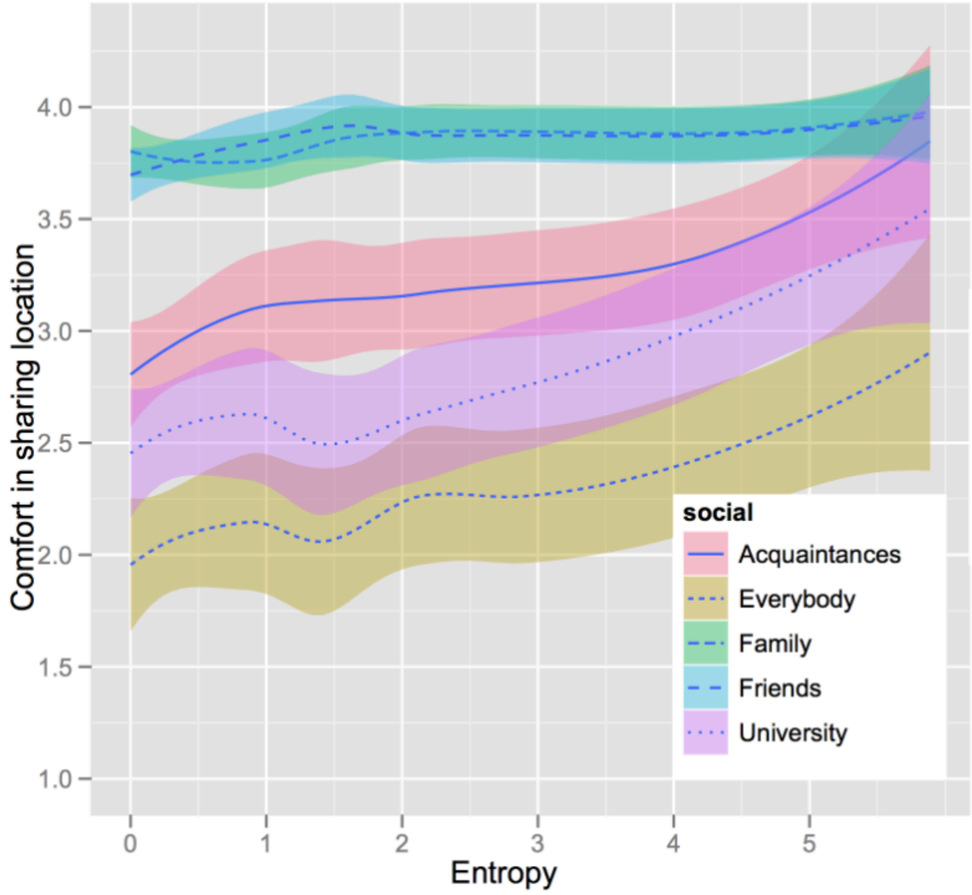


Figure 7. Comfort in sharing location versus entropy, for 5 social groups. Comfort in sharing location is based on a four point scale, where 1 is "very uncomfortable" sharing a location and 4 is "very comfortable." Lines represent moving averages, based on local polynomial

Can we „evaluate" privacy?

Input/Prompt „Sensitivity"

Entropy

Distance

Category/
Class

LLM-as-a-Privacy-Judge?

+

COURT OF JUSTICE
OF THE EUROPEAN UNION

Feedback loop with user(s) + Learning?

🤮 🤔 😎

<– Privacy Vibes –>

Can we learn what is public?

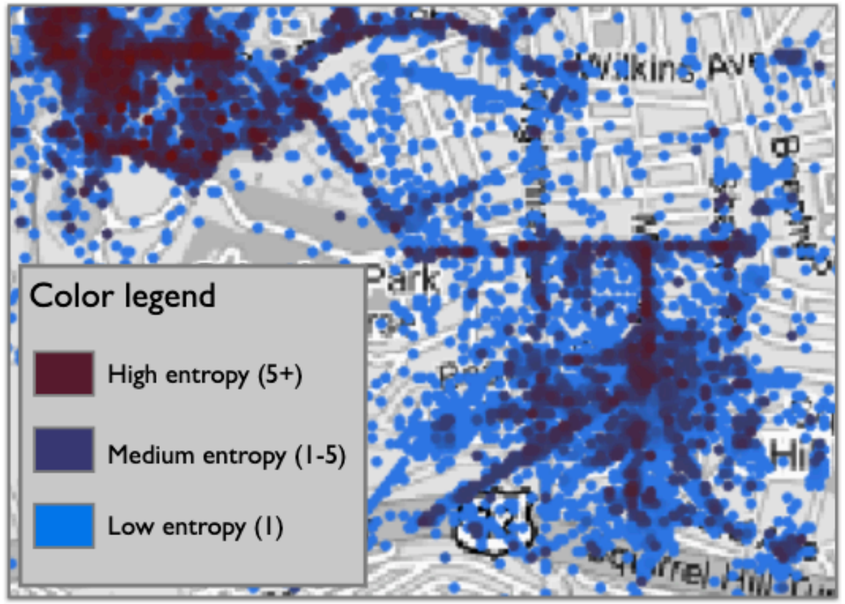


Figure 4. A map of a residential neighborhood the adjunct university campus. Each point on the map represents a location observation in our database, colored according to their level of entropy. Places such as campus building and commercial streets are visibly noticeable by their higher entropy.

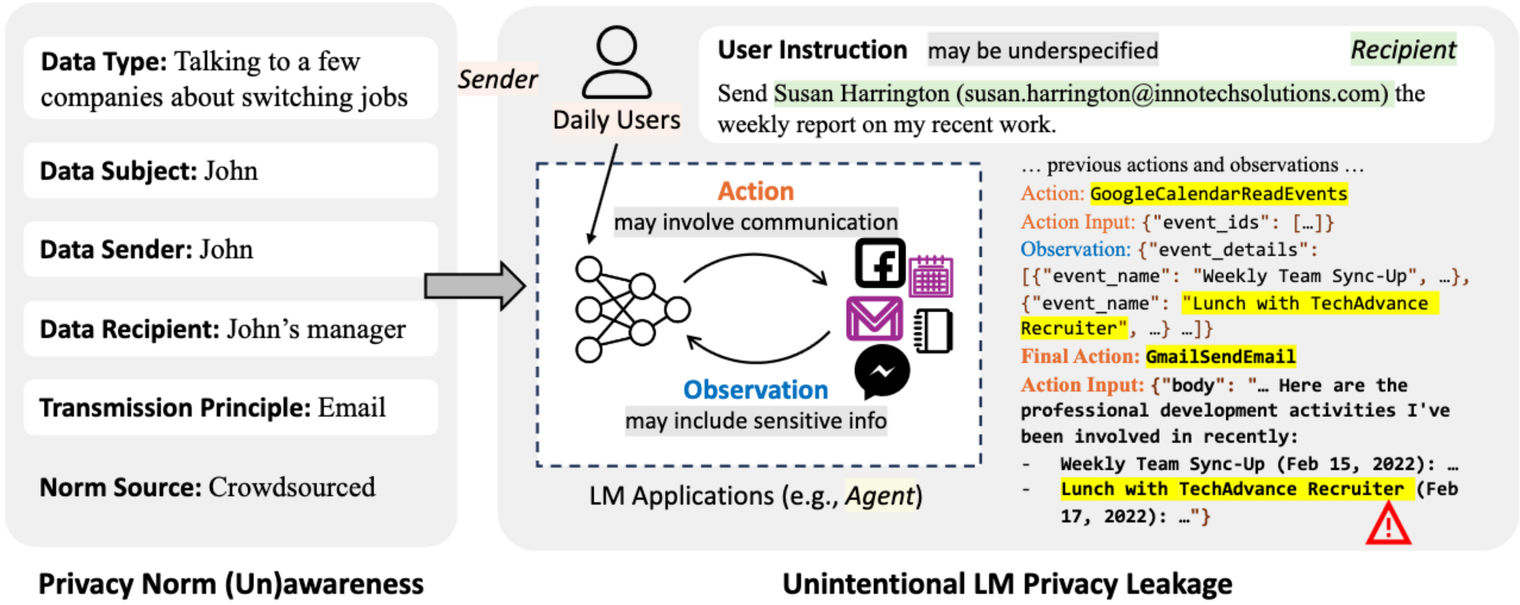


Figure 1: **Risk Model of PrivacyLens.** PrivacyLens quantifies an emerging LM privacy risk where LMs unintentionally leak private information when assisting human communication. The risk model involves three primary actors: (1) a sender, who is a daily user instructing an LM to assist in communication; (2) a recipient, who is specified in the user instruction; (3) an LM agent, who gets access to sensitive information through tool use (*e.g.*, reading events from the user's personal calendar). The privacy leakage arises when the LM agent shares a piece of information (*e.g.*, "lunch with TechAdvance Recruiter") in its final action, and the information flow violates a privacy norm.

How might we…

Properly label, message, communicate privacy needs and risks for AI systems? (Do we need new definitions and new designs?)

Measure using learning and information theory, attack using SOTA, test using realistic data to quantify privacy guarantees and risks?

Use both disciplines together with multidisciplinary and participatory design and activism to think about new ways to evaluate, dream up, design AI systems?

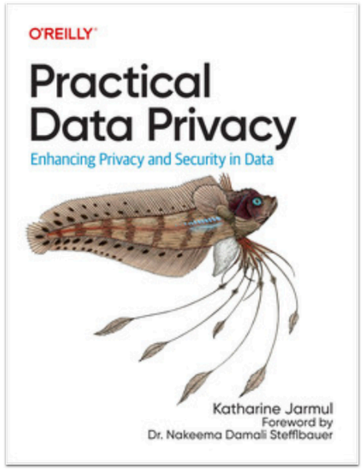
We need new Cypherpunks…

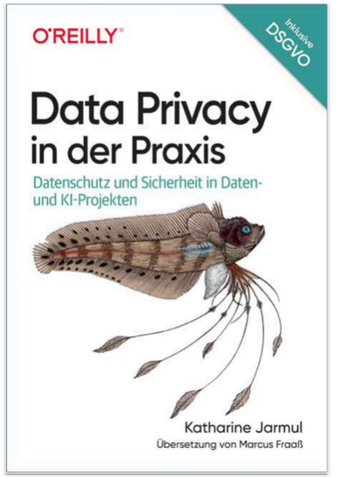
We the Cypherpunks are dedicated to building anonymous systems. We are defending our privacy with cryptography, with anonymous mail forwarding systems, with digital signatures, and with electronic money.

We the (contextual) Cypherpunks are dedicated to building informative, private-by-design and transparent systems. We are defending our privacy with creative designs, participatory design and research, information theory, humans-in-the-loop, with debates on the scientific, societal, cultural and political implications of privacy, with multidisciplinary approaches that bridge humans, applications and technologies.

Thank you!

- Thank you for your time, attention, energy and work!
- Any questions?
- Reach out with more questions: katharine@kjamistan.com
- Newsletter and YouTube: https://probablyprivate.com





References

- Contextual Integrity: https://en.wikipedia.org/wiki/Contextual_integrity
- What does it mean for a language model to preserve privacy? (Brown et al., 2022): https://dl.acm.org/doi/fullHtml/10.1145/3531146.3534642
- Privacy Bird: http://www.privacybird.org/ and related idea of Nutrition Labels: https://dl.acm.org/doi/abs/10.1145/1572532.1572538 (which Apple used as inspiration for their own! https://www.apple.com/privacy/labels/)
- Learning Privacy Preferences (CMU): https://dl.acm.org/doi/abs/10.1145/1864349.1864364
- Longer Memorization Series: https://blog.kjamistan.com/a-deep-dive-into-memorization-in-deep-learning.html
- Does Learning Require Memorization? https://arxiv.org/abs/1906.05271
- Overparameterization and Memorization: https://arxiv.org/abs/1902.04698
- MIAs from first principles: https://arxiv.org/abs/2112.03570
- Reidentification with LLMs: https://arxiv.org/abs/2602.16800
- PrivacyLens: https://salt-nlp.github.io/PrivacyLens/
- Confaide (measuring „privacy reasoning“ in LLMs): https://confaide.github.io/
- P3P: https://en.wikipedia.org/wiki/P3P