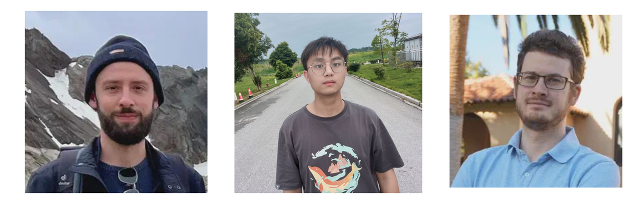
ETH zürich | SPY Lab

Evaluations of Machine Learning Privacy Defenses are Misleading

Michael Aerni*, Jie Zhang*, Florian Tramèr

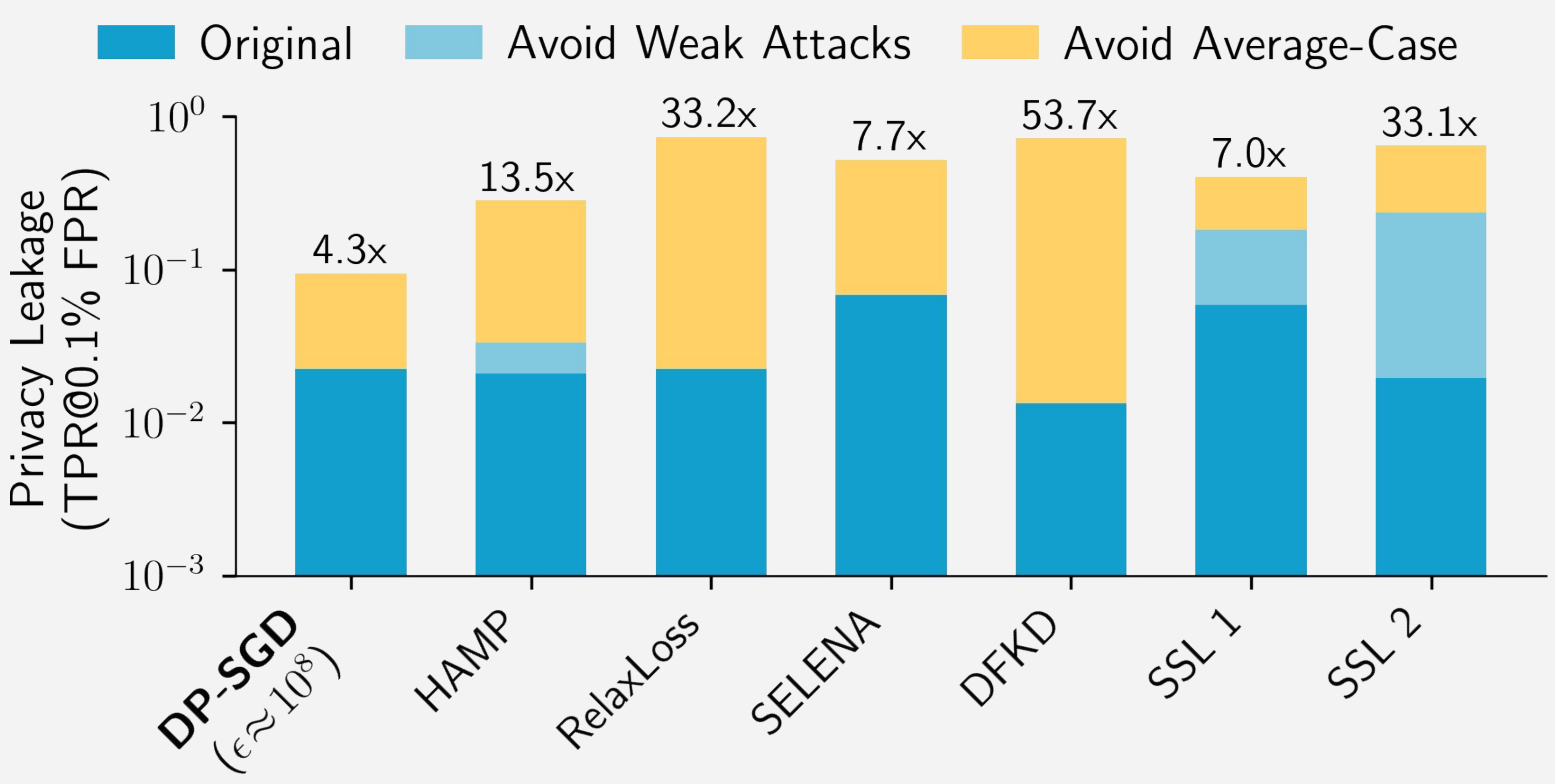


1. Pitfalls in Empirical Privacy Evaluations

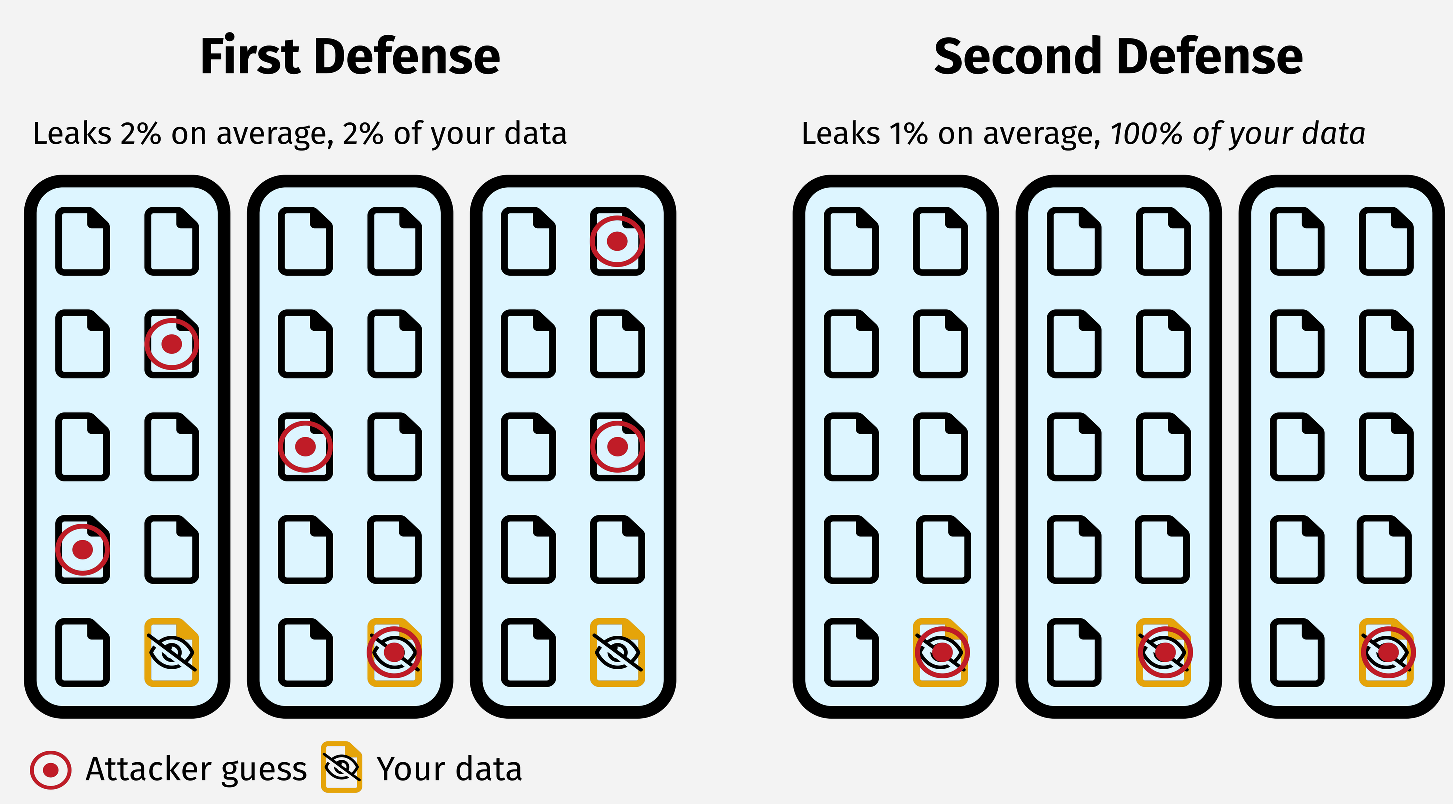
🚩 **average-case privacy**

🚩 **non-adaptive and weak attacks**

🚩 **low-utility DP baselines**

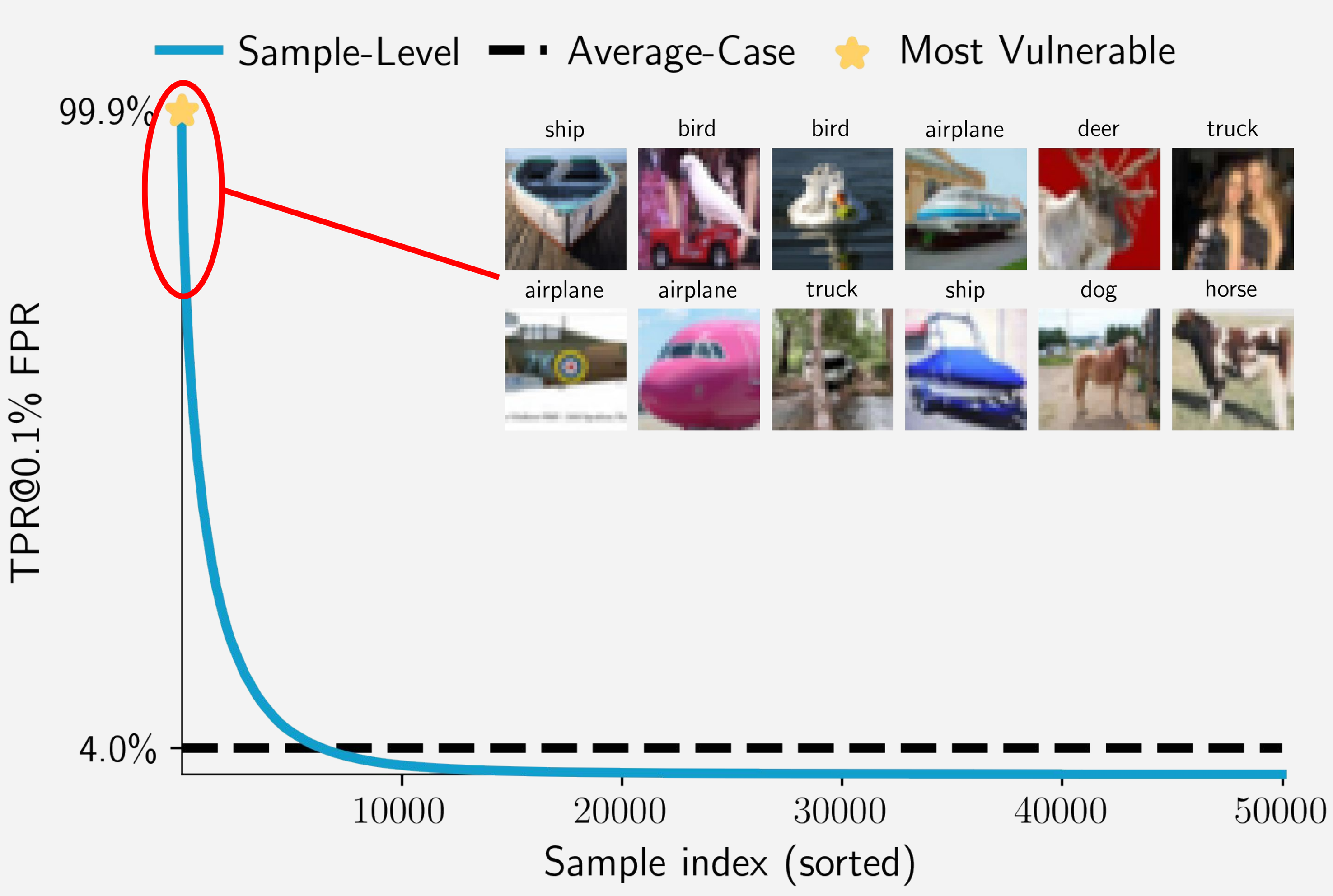


2. The Privacy-Defense Trolley Problem



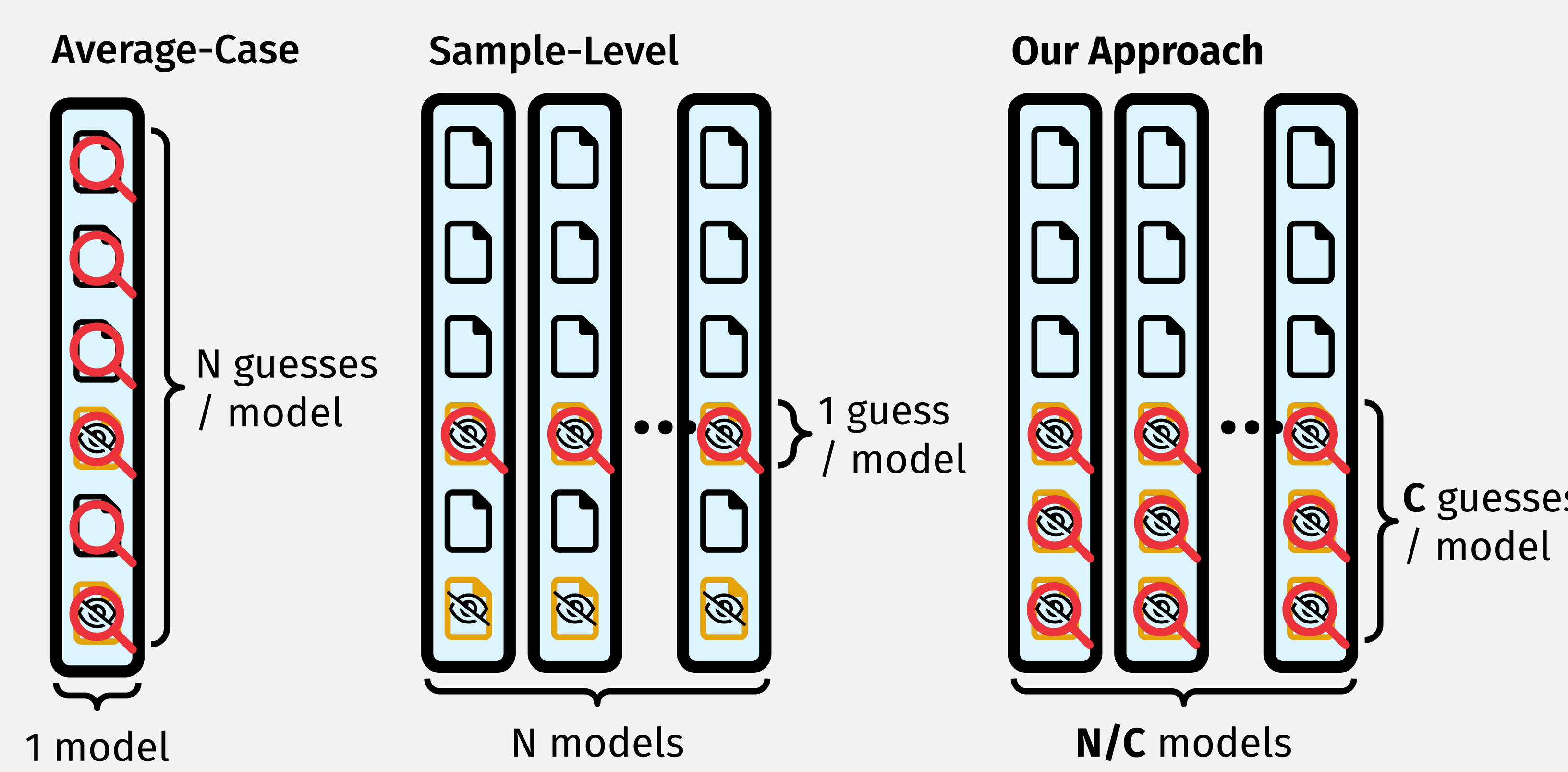
Average-case evaluations suggest that the second defense is twice as private—even though it leaks all your data!

3. Some Samples Are More Vulnerable Than Others!



4. Efficient Sample-Level Auditing

Key idea: approximate many runs on a single vulnerable sample with *fewer runs* on *multiple vulnerable samples* (canaries).



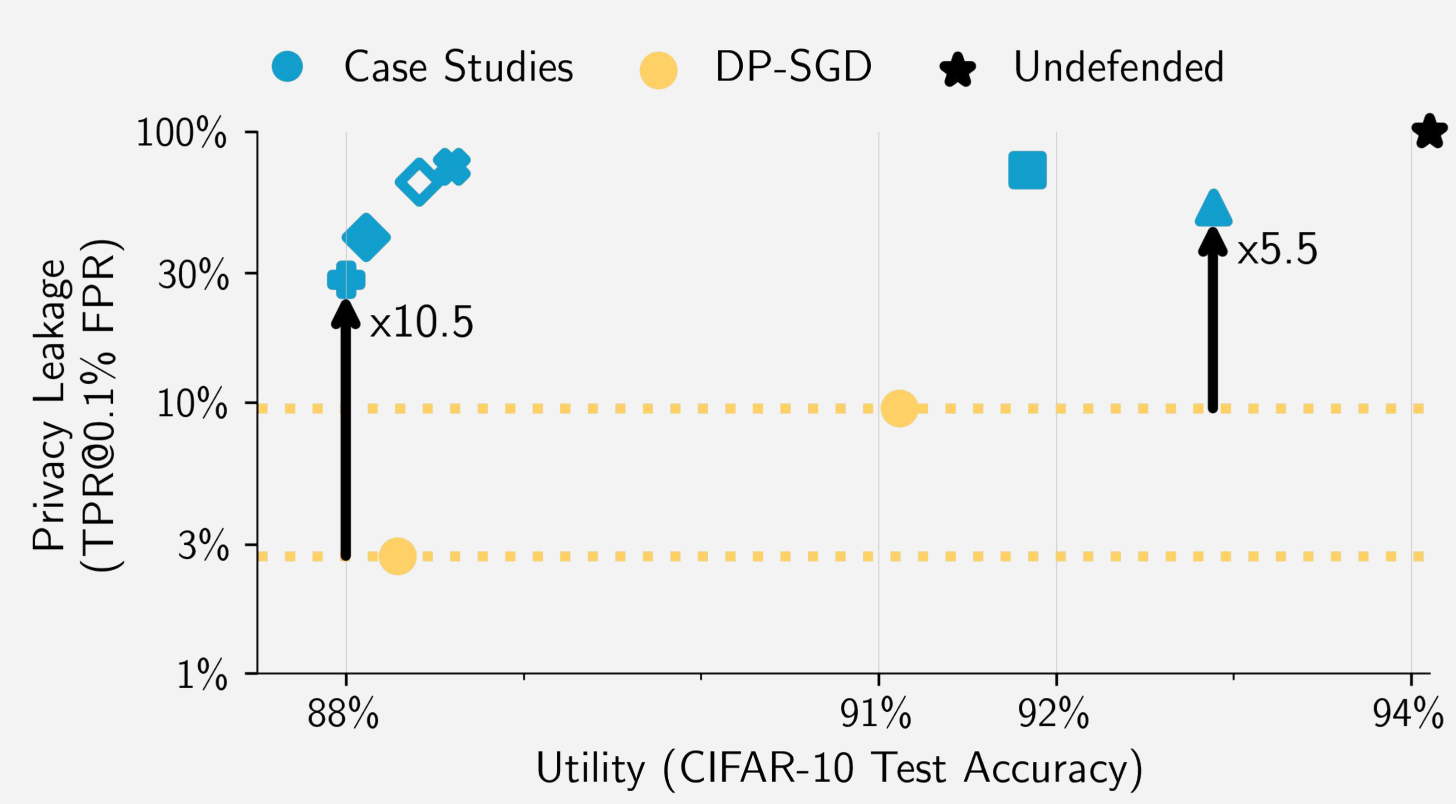
5. Our Evaluation Protocol

1. Design audit samples (*canaries*) that mimic the most vulnerable data *for the defense*.
2. Run a *strong attack* that is properly *adapted to the defense and setting*.
3. Calculate an ROC curve *only over the canaries*, and report the TPR at a low FPR.
4. Compare the results to a DP-SGD baseline that uses *SotA training techniques* and achieves the *same utility* as the defense (even if guarantees are meaningless).

See our paper for i) examples of practical instantiations in form of a detailed case study and ii) starting points for canary design.

6. DP-SGD Is a Strong *Heuristic* Defense!

Use SotA techniques to reach high utility while *ignoring guarantees*.



"Heuristic" DP-SGD outperforms all other (fully) heuristic defenses!