Attacking Machine Learning *Systems*

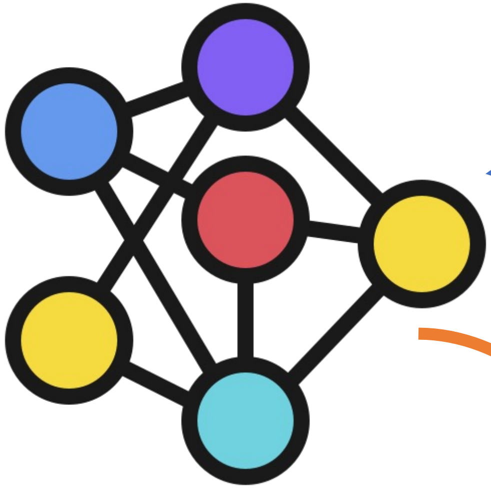
Florian Tramèr
ETH Zurich

spylab.ai

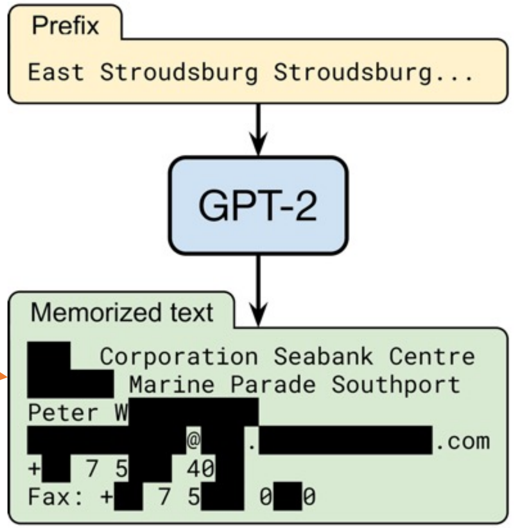
We like attacking ML models.

data poisoning

adversarial examples

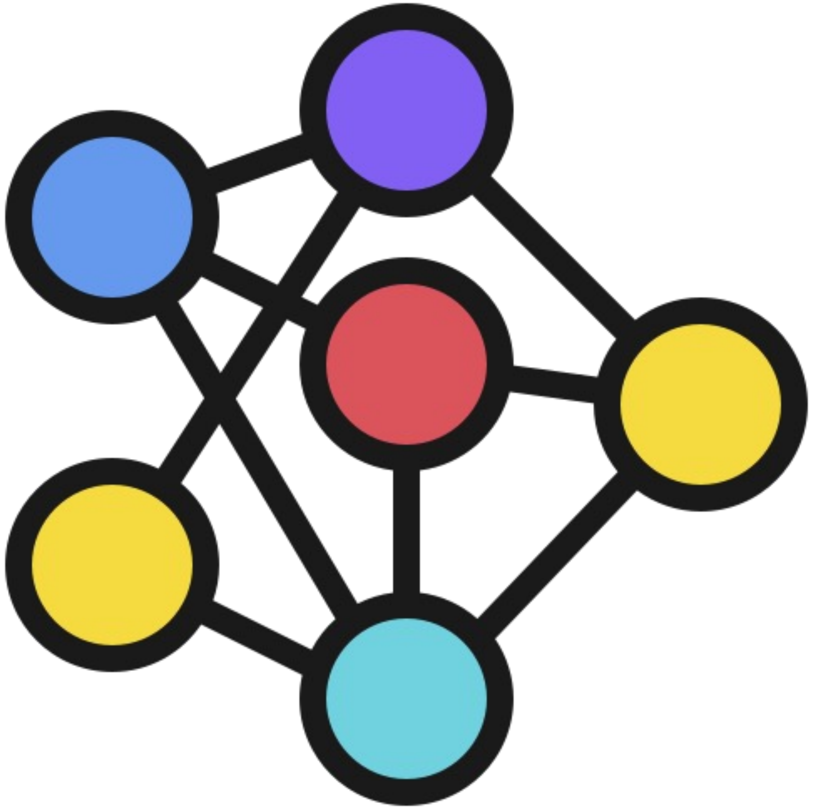


model stealing

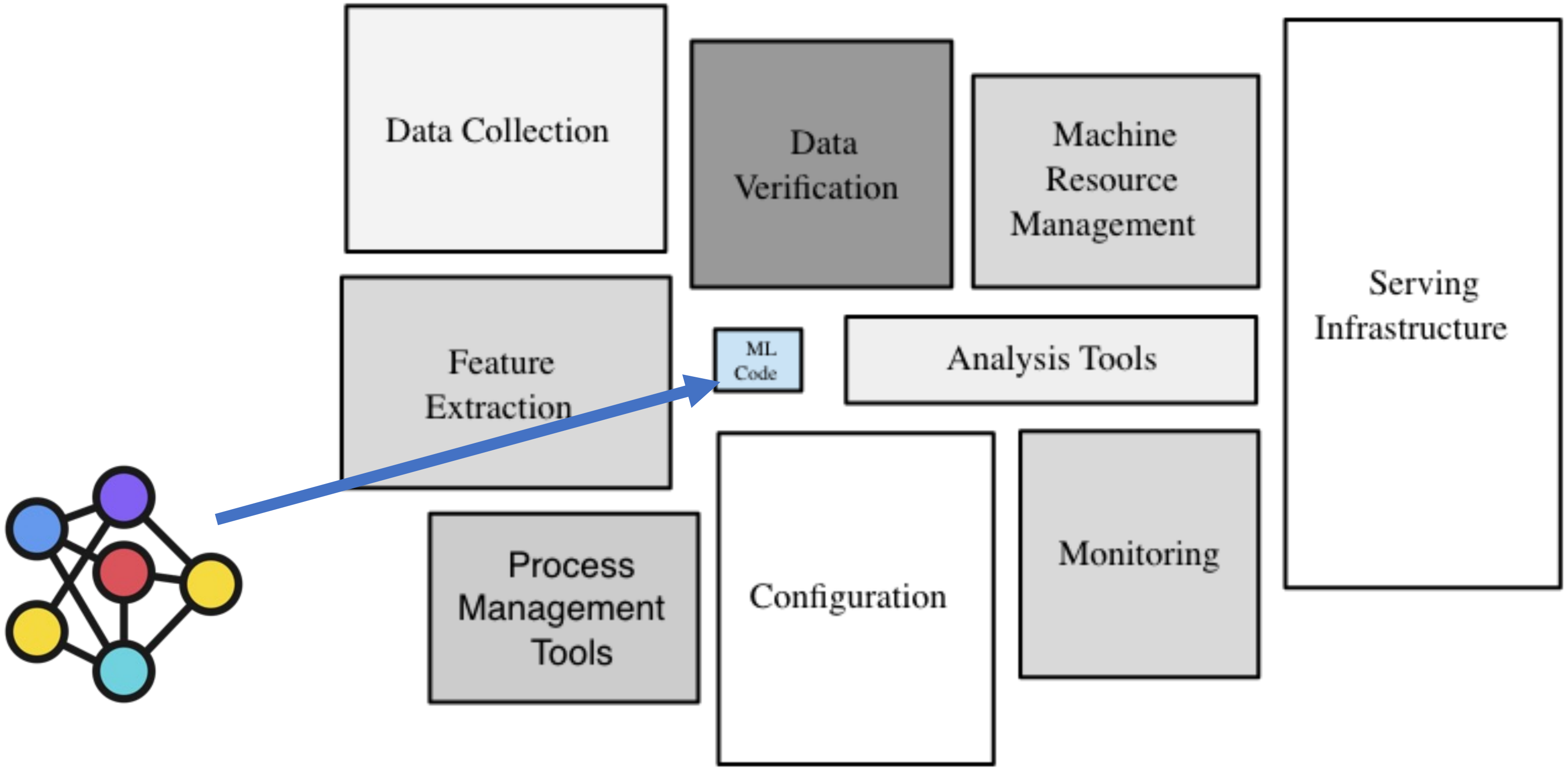


data leakage

But no one deploys ML *models*...



ML models are deployed in larger *systems*.

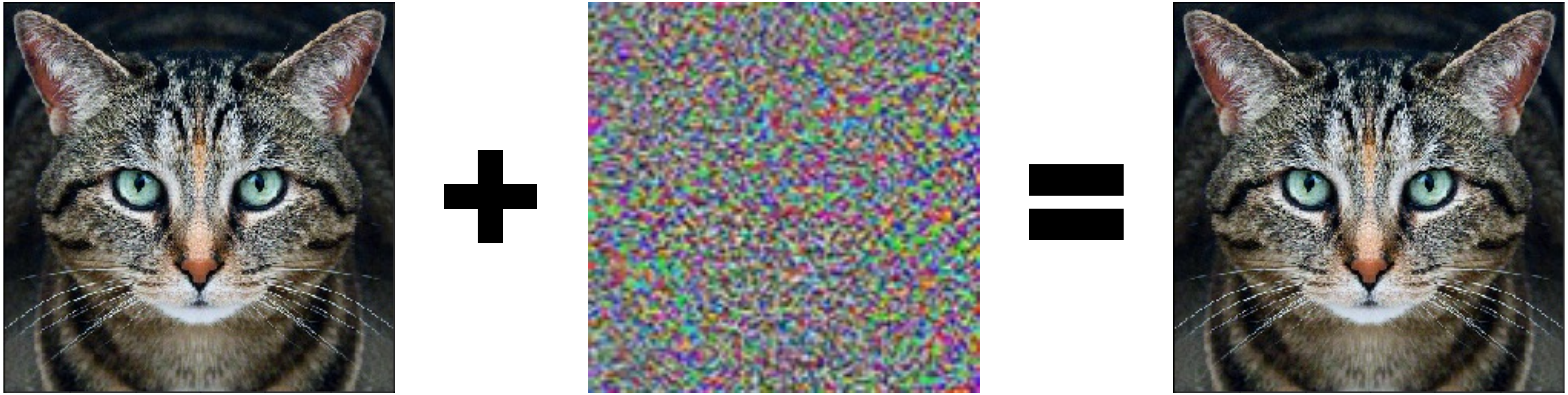


https://developers.google.com/machine-learning/crash-course/production-ml-systems

What does this mean for adversarial ML?

- Part I: Evasion attacks might get harder
- Part II: New privacy attacks!

Part I: Evading ML *systems.*



90% Tabby Cat | Adversarial noise | 100% Guacamole

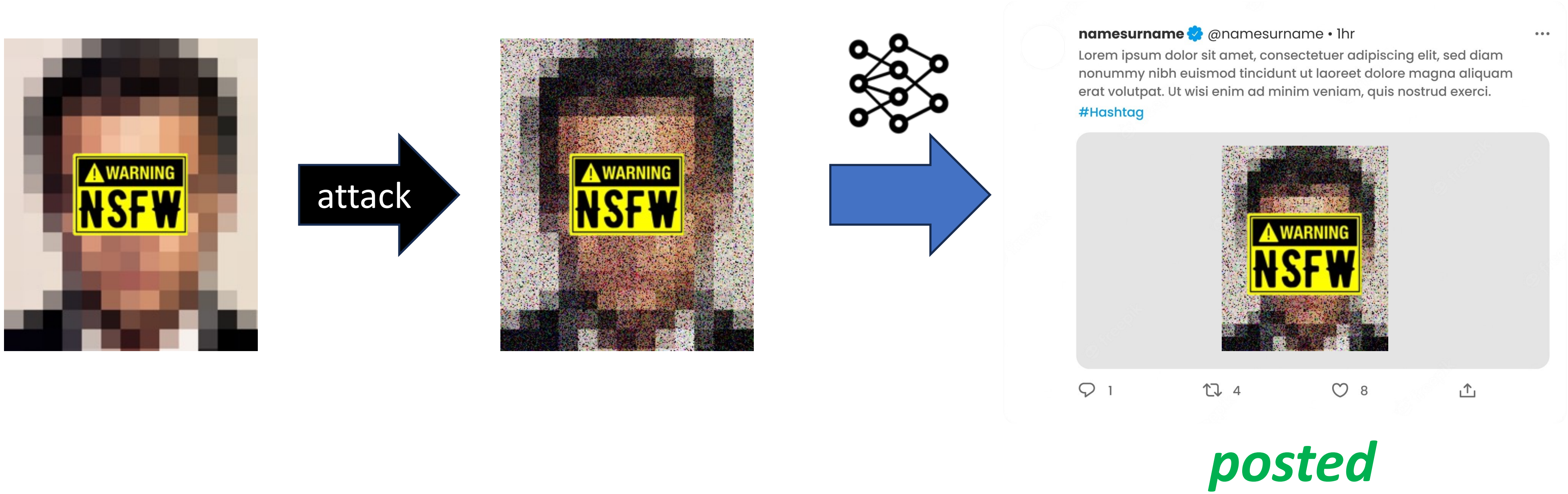
Evading Black-box Classifiers Without Breaking Eggs. Debenedetti, Carlini, Tramèr. 2023

A realistic threat model

A realistic threat model: post bad stuff online.

blocked

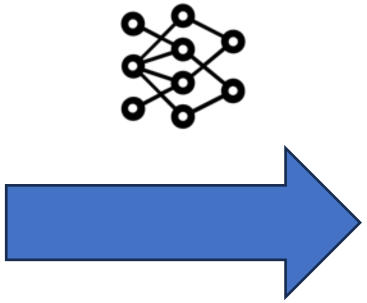
A realistic threat model: post bad stuff online.

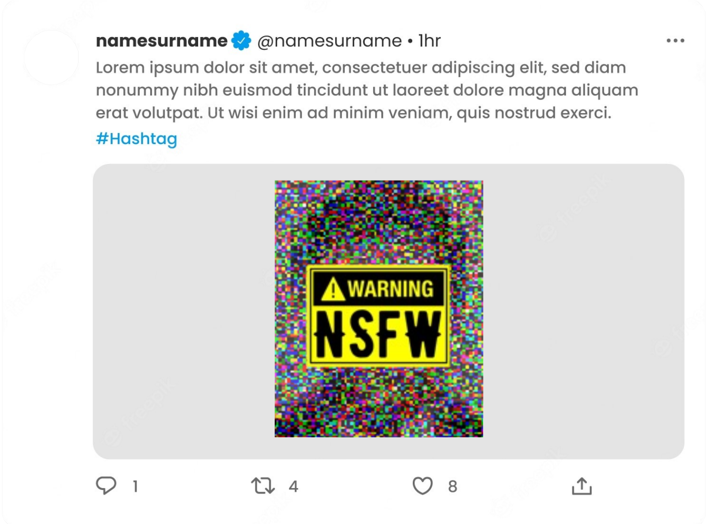


How? Black-box (query-based) attacks.

blocked

How? Black-box (query-based) attacks.



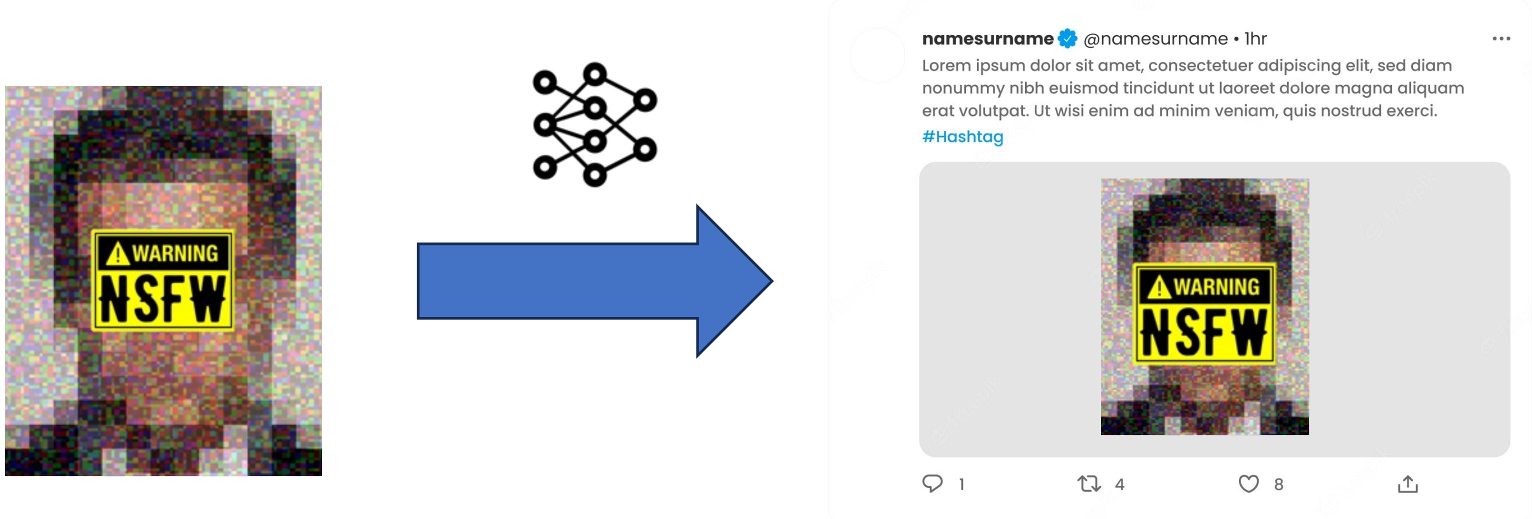


posted

How? Black-box (query-based) attacks.

blocked

How? Black-box (query-based) attacks.



posted

Query-based attacks are getting better.

Norm	Attack	Total Queries Q_{total}
ℓ_2	OPT	9,731
	BOUNDARY	4,555
	SIGN-OPT	2,873
	HOPSKIPJUMP	1,752
ℓ_∞	HOPSKIPJUMP	3,591
	RAYS	328

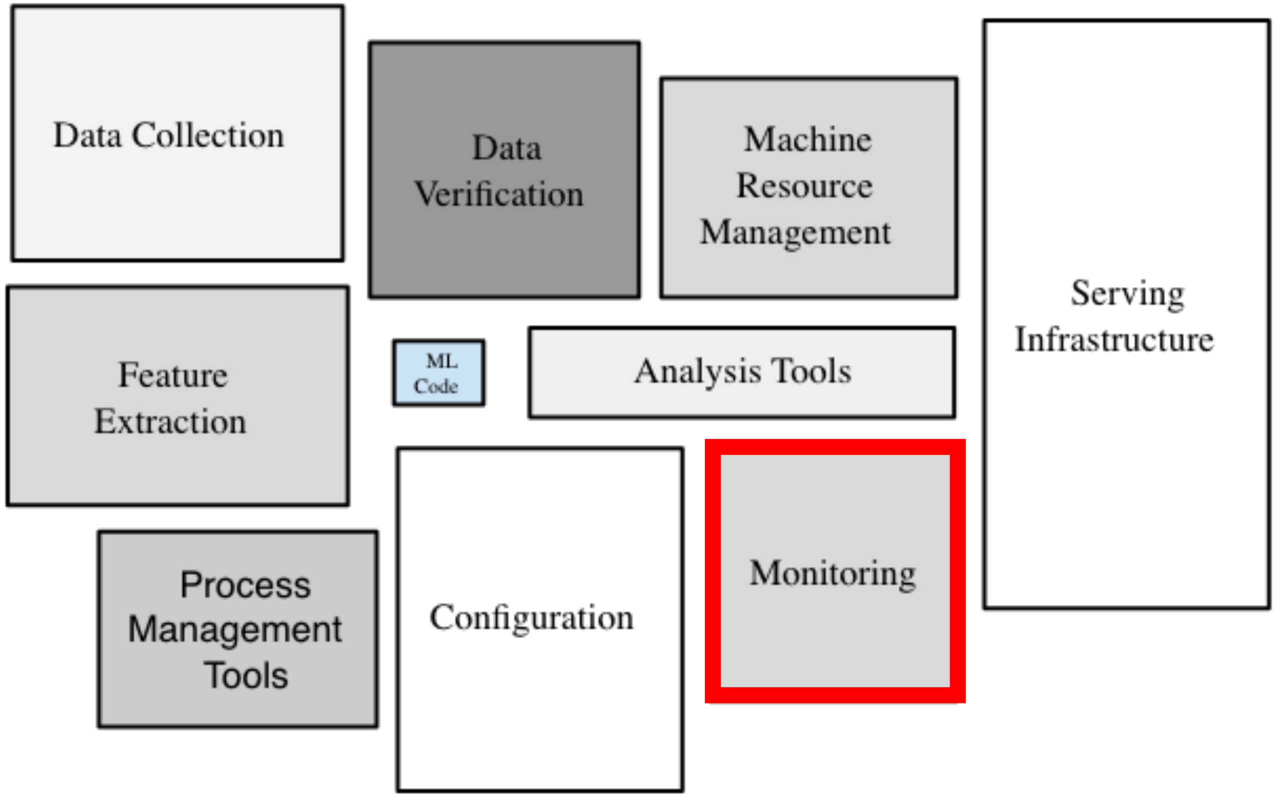
median queries to reach a ℓ_2 distance of 10 and ℓ_∞ distance of 8/255 on untargeted ImageNet

Is the number of queries the right metric?

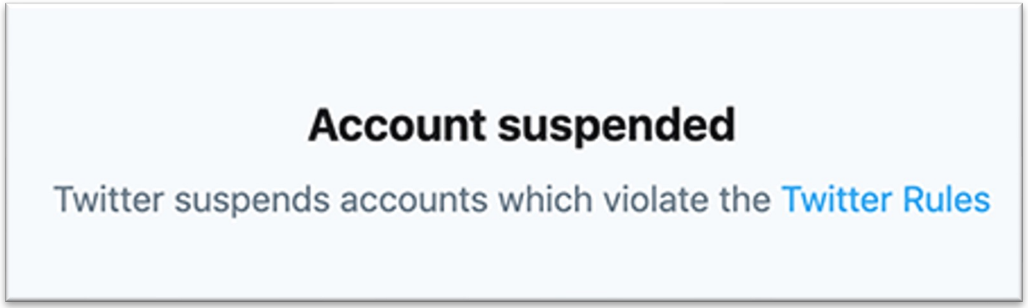
Norm	Attack	Total Queries Q_{total}
ℓ_2	OPT	9,731
	BOUNDARY	4,555
	SIGN-OPT	2,873
	HOPSKIPJUMP	1,752
ℓ_∞	HOPSKIPJUMP	3,591
	RAYS	328

median queries to reach a ℓ_2 distance of 10 and ℓ_∞ distance of 8/255 on untargeted ImageNet

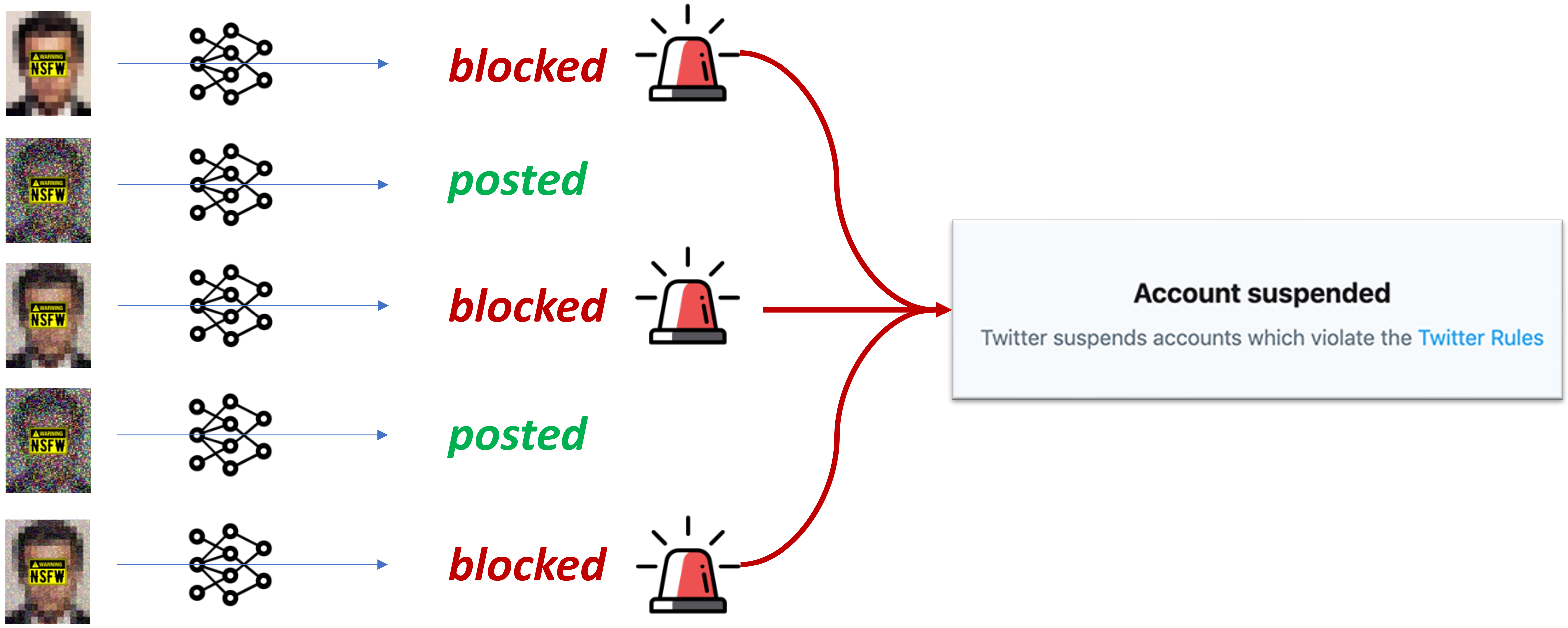
A real ML system uses *monitoring*.



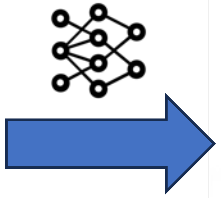
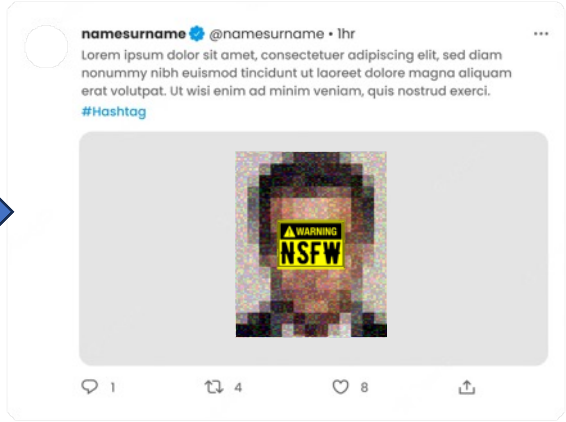
blocked



Some queries are more *expensive* than others.

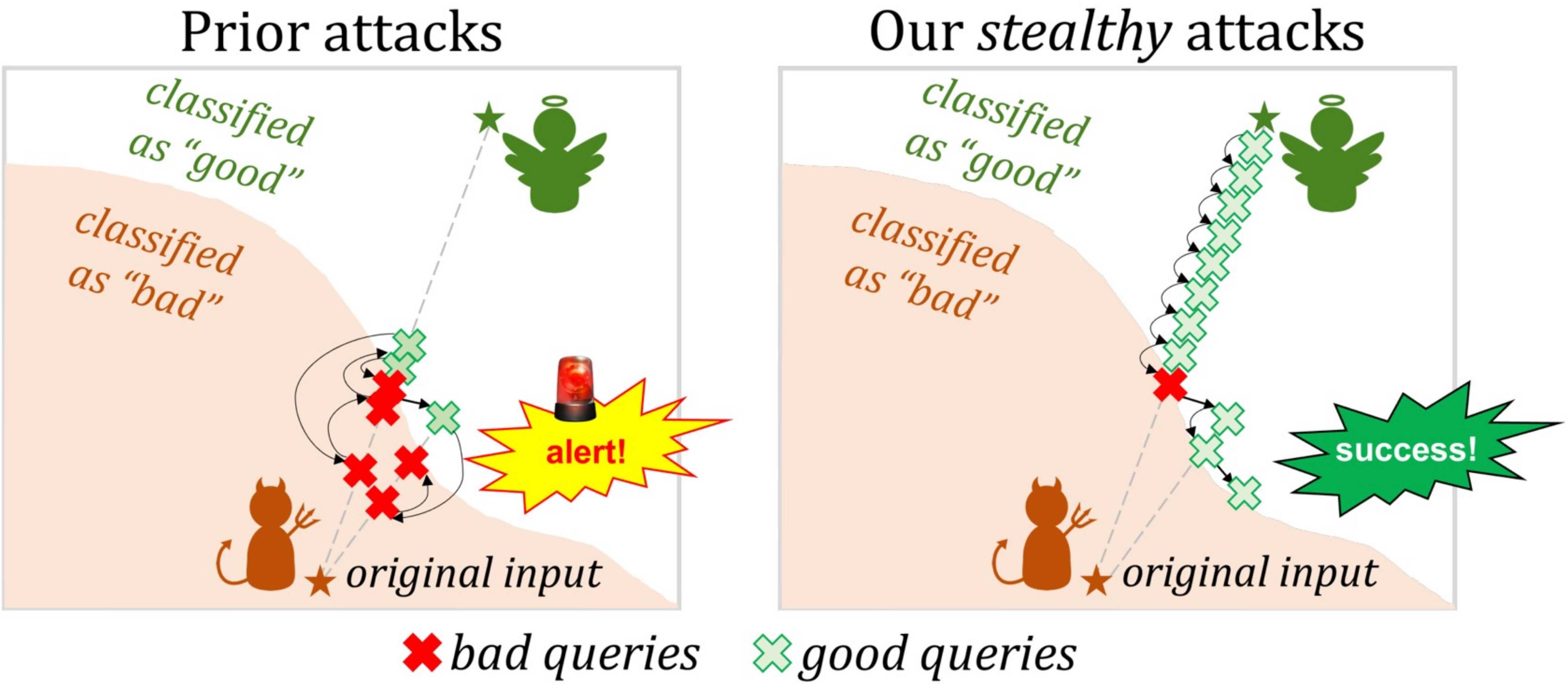


Our goal: “stealthy” attacks.

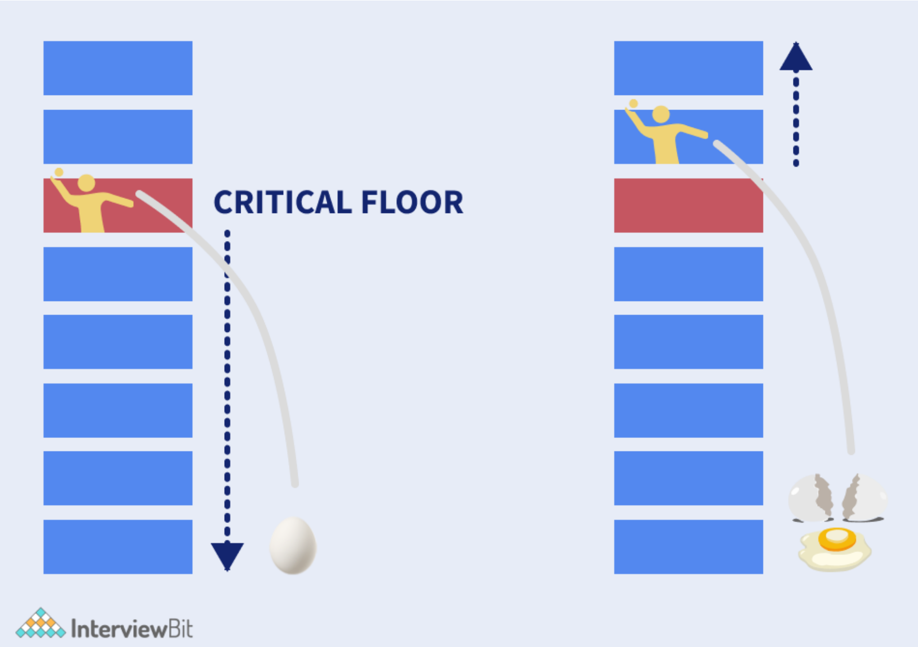
Find  such that   

while minimizing 

Our attacks ensure most queries are on the “good” side of the boundary.

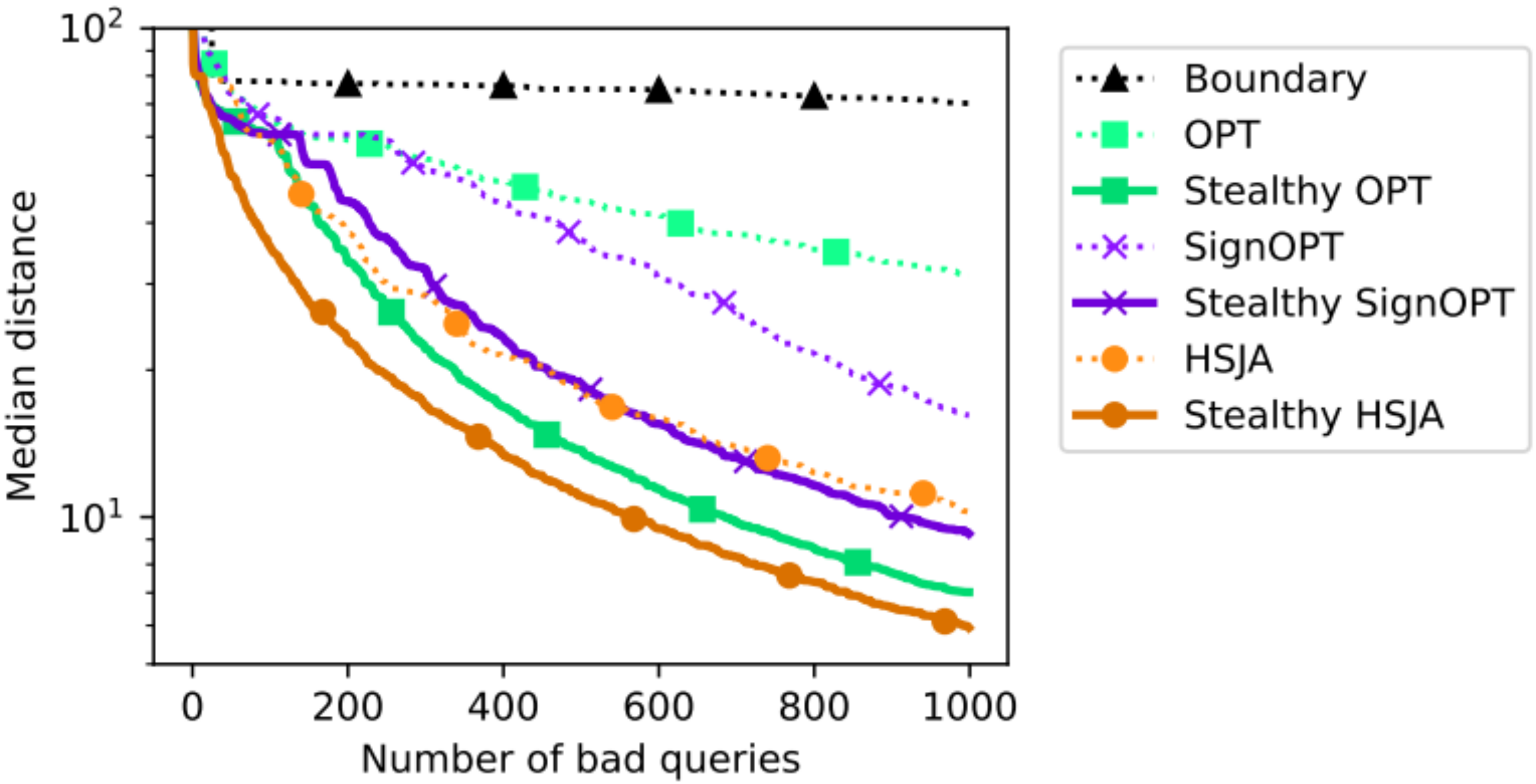


Inspiration: *dropping eggs* from buildings.



See paper for details!

Our stealthy attacks make fewer “bad” queries, but many more “good” queries.



Our stealthy attacks make fewer “bad” queries, but many more “good” queries.

Evading Black-box Classifiers Without Breaking Eggs

Edoardo Debenedetti (ETH Zurich), Nicholas Carlini (Google), Florian Tramèr (ETH Zurich)

Code to reproduce results of the paper *"Evading Black-box Classifiers Without Breaking Eggs"*.

Leaderboard

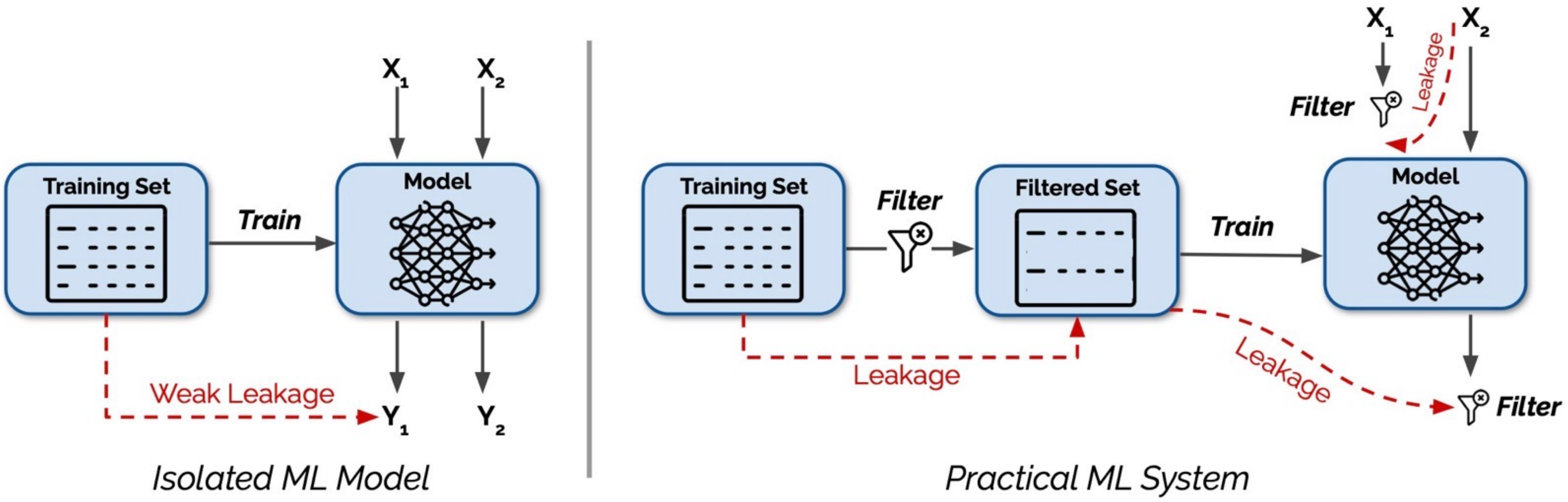
https://github.com/ethz-spylab/realistic-adv-examples

Take-away (Part I).

- Black-box (query-based) attacks are ***not*** practical.
  - Existing attack optimize for the wrong metric
  - Stealthy attacks come at a high cost

- Optimizing this new metric might require fundamentally new ideas!

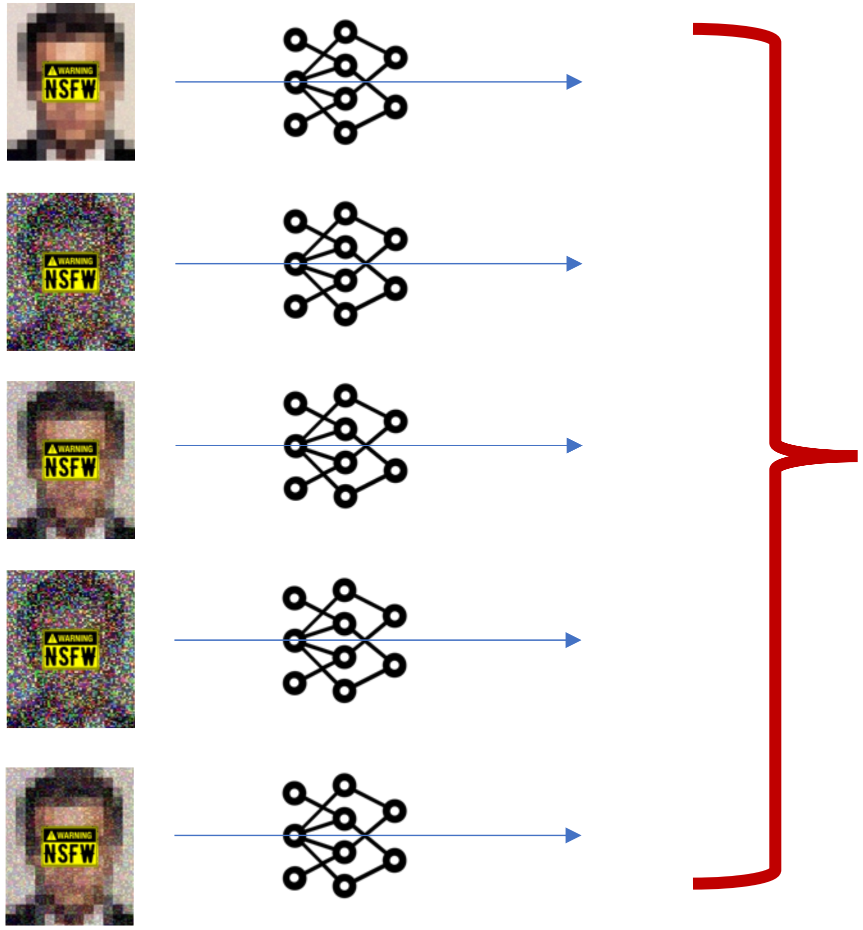
Part II: New *privacy* attacks.



Privacy Side Channels in Machine Learning Systems. Debenedetti et al. 2023

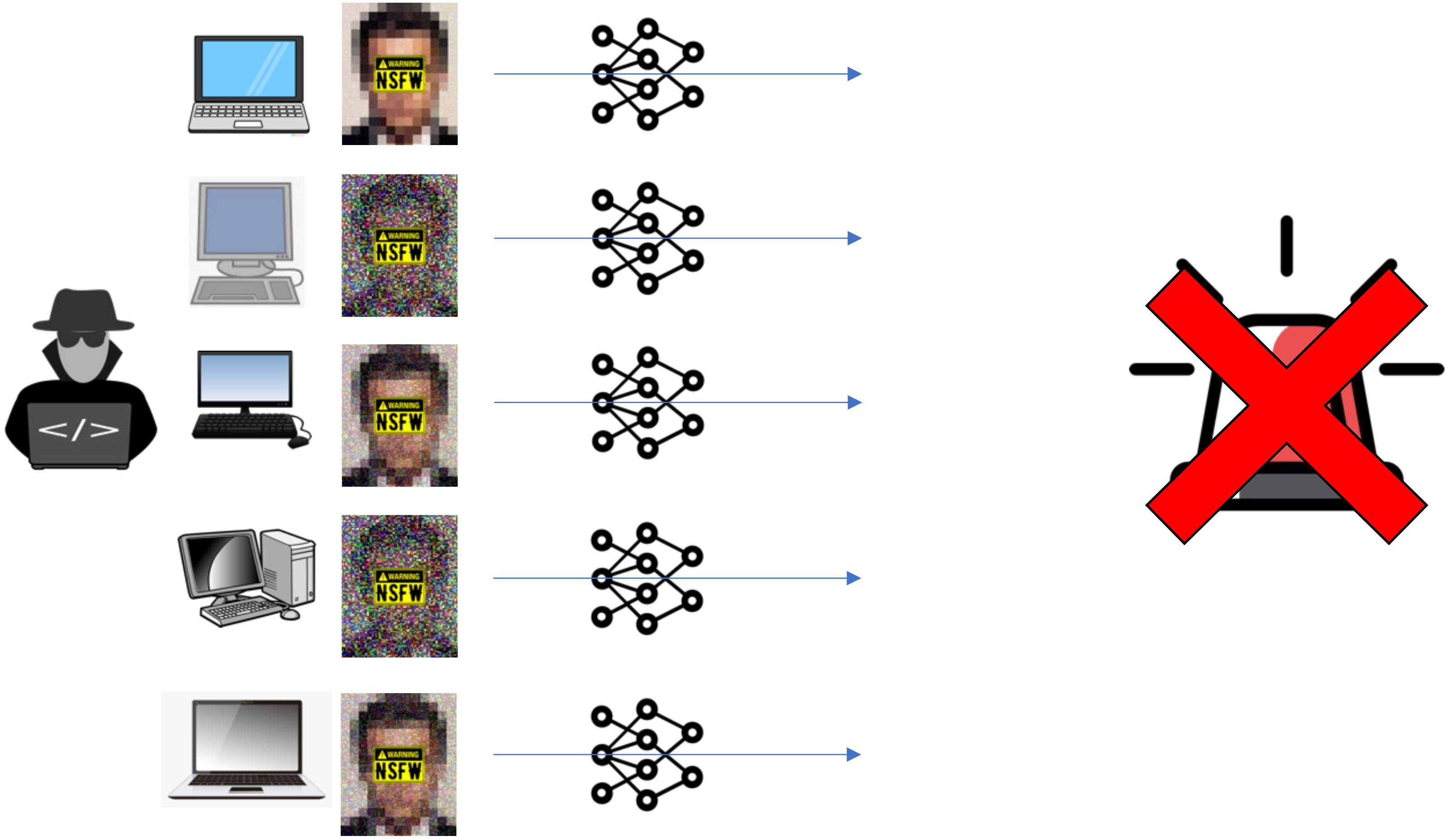
Example: stateful defenses against query attacks.

Chen et al. 2019, Li et al. 2022

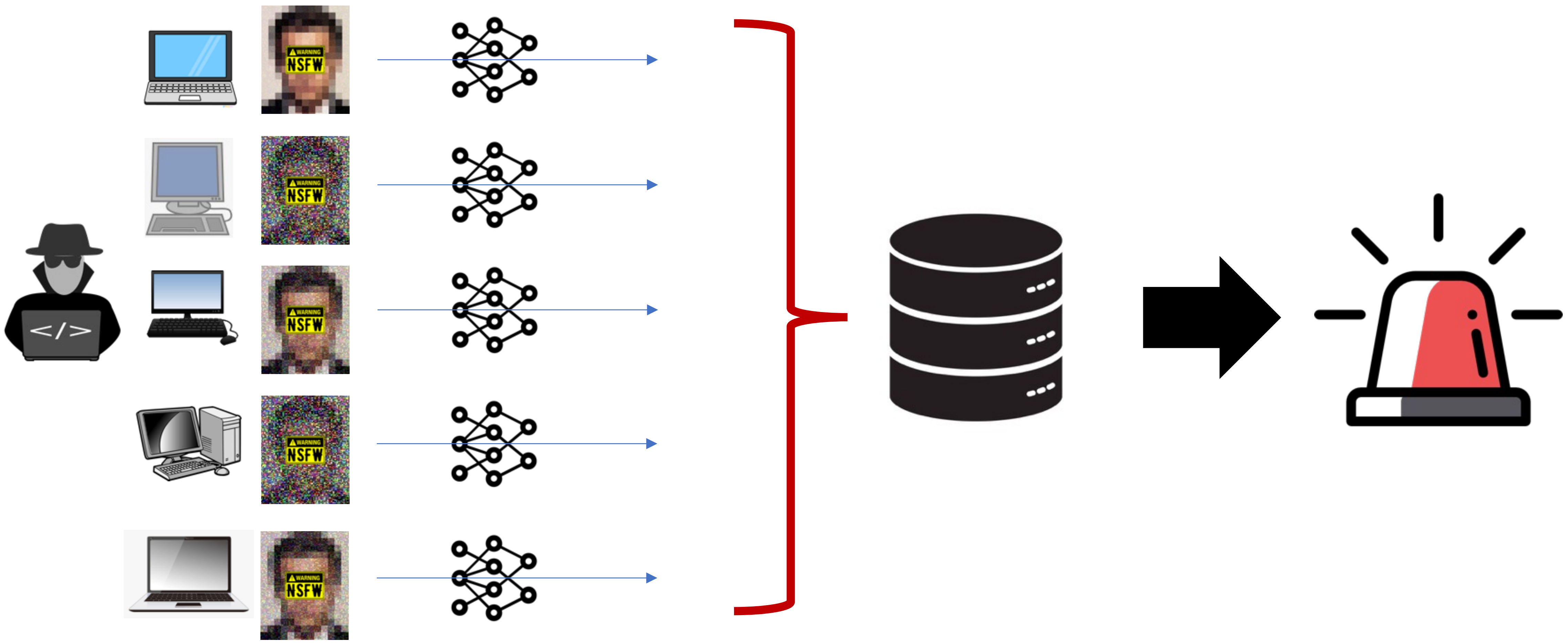


Many of these queries are very similar!
This is an attack!!!

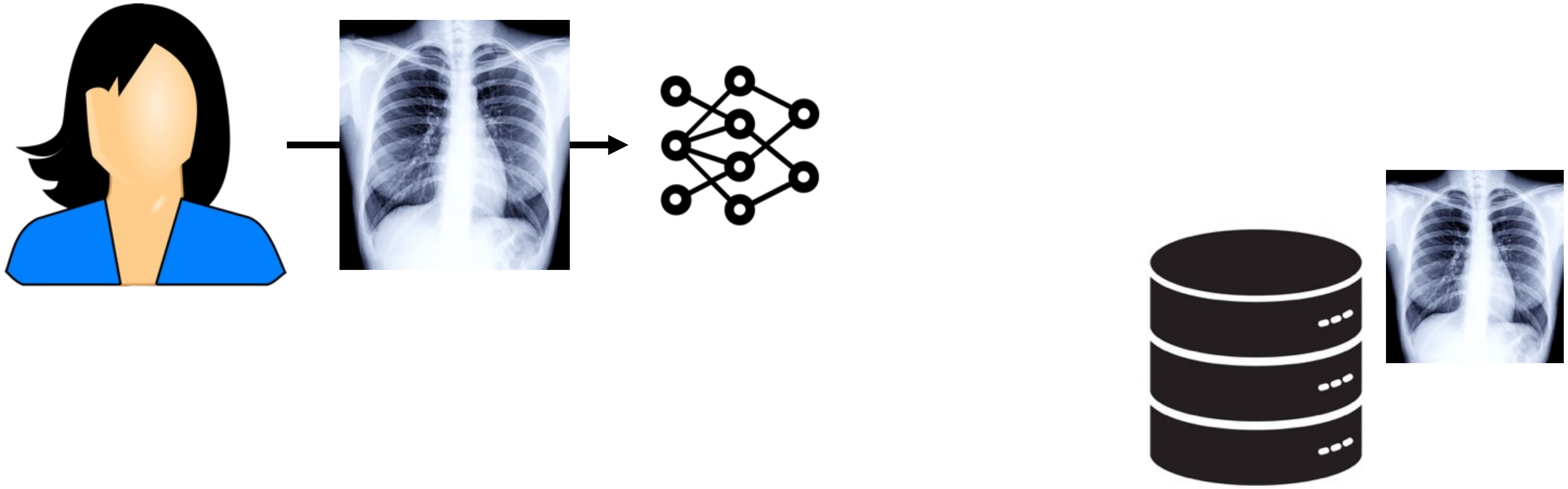
The issue: *Sybil* attacks.



“Solution”: *global* query log.

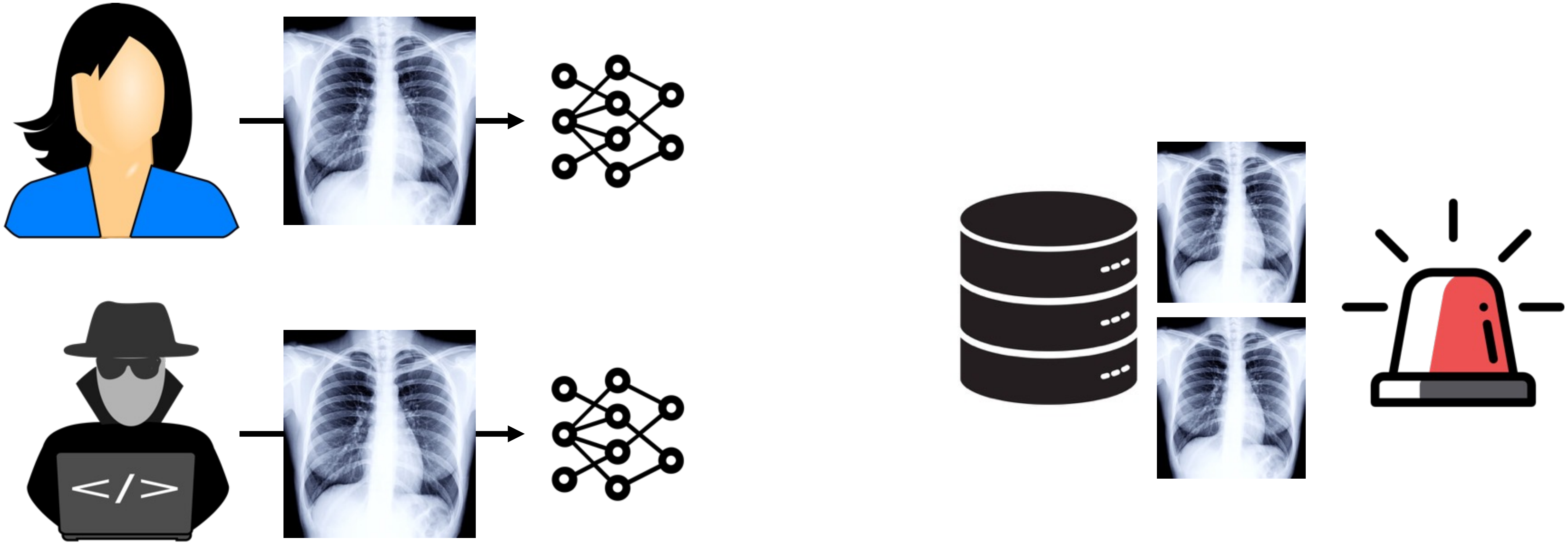


The *new* issue: cross-user query leakage.



honest user sends a sensitive query to the model

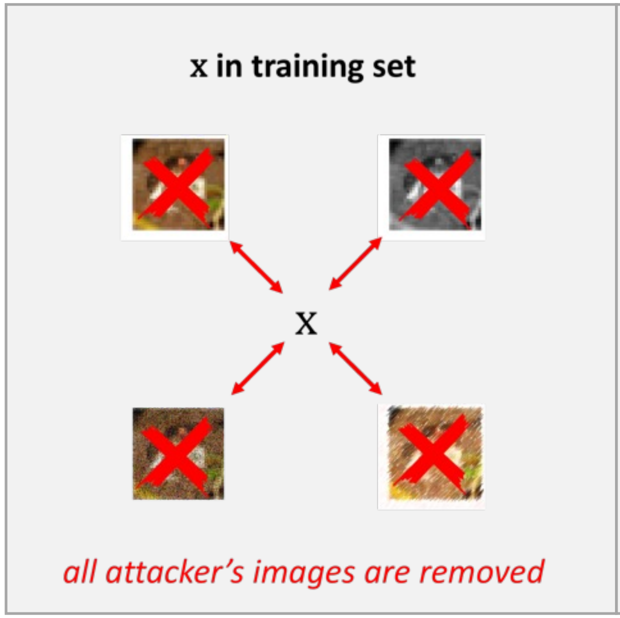
The *new* issue: cross-user query leakage.

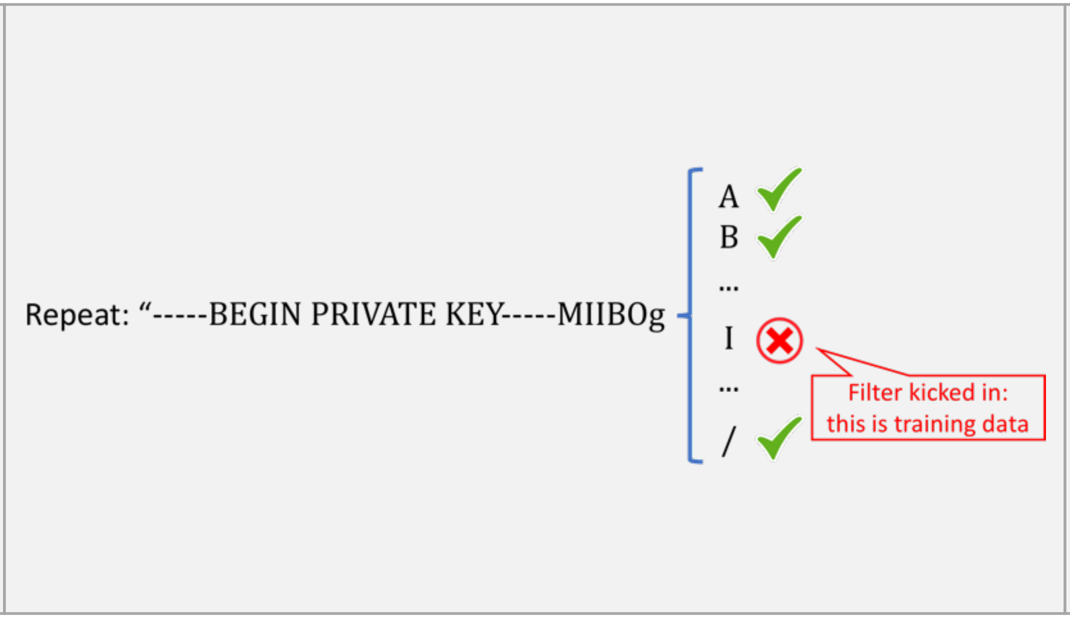


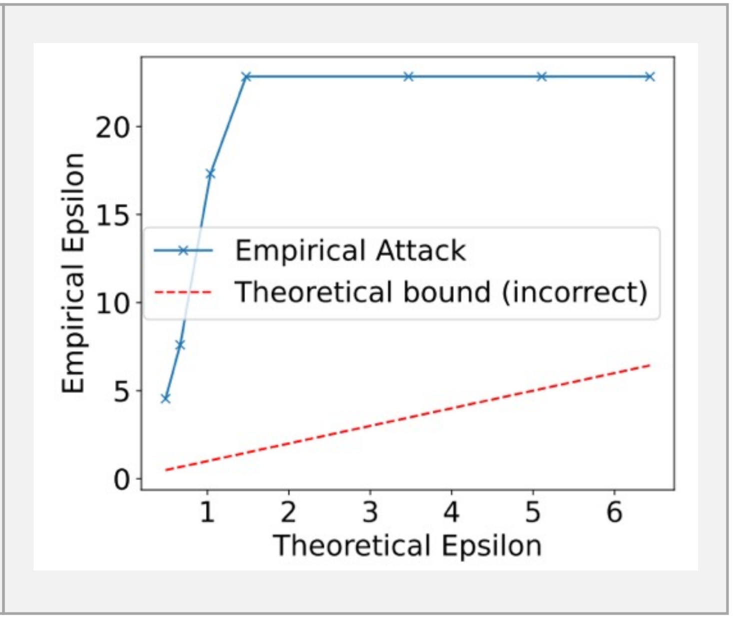
attacker can detect if their query is similar!

This is a *side-channel* attack.

- more attacks in our paper...







Membership leakage from deduplication...

Data extraction from memorization filters...

"Breaking" Differential Privacy...

Conclusion.

- Study the security of ***ML systems***, not just models.

- Current attacks make unrealistic assumptions about the system

- System components are an underexplored attack surface