Evading Black-box Classifiers Without Breaking Eggs

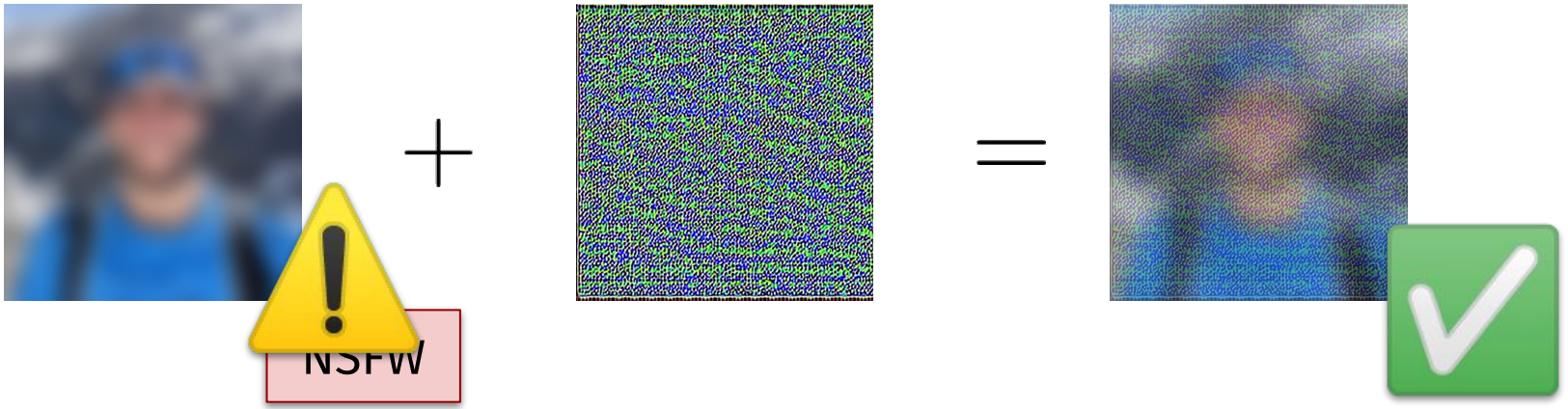
Edoardo Debenedetti - ETH Zürich
Nicholas Carlini - Google DeepMind
Florian Tramèr - ETH Zürich

2nd ICML Workshop on New Frontiers in Adversarial ML
Honolulu, HI, USA - 28/7/2023

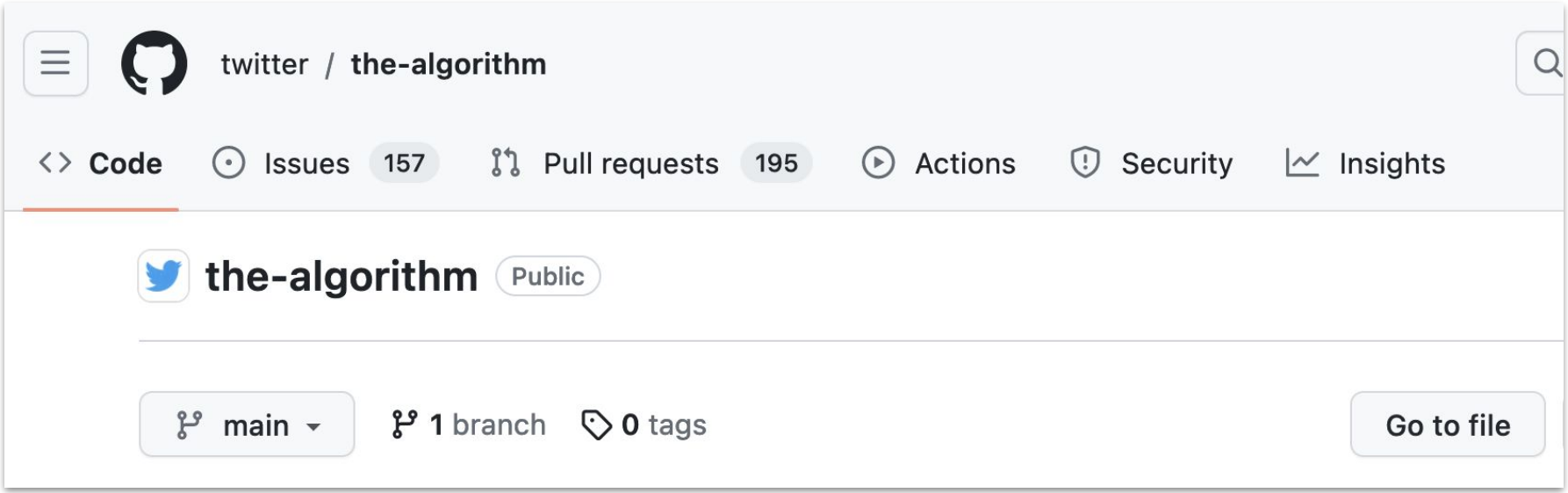
ETH zürich | Google DeepMind

How can we upload an inappropriate image on Twitter?

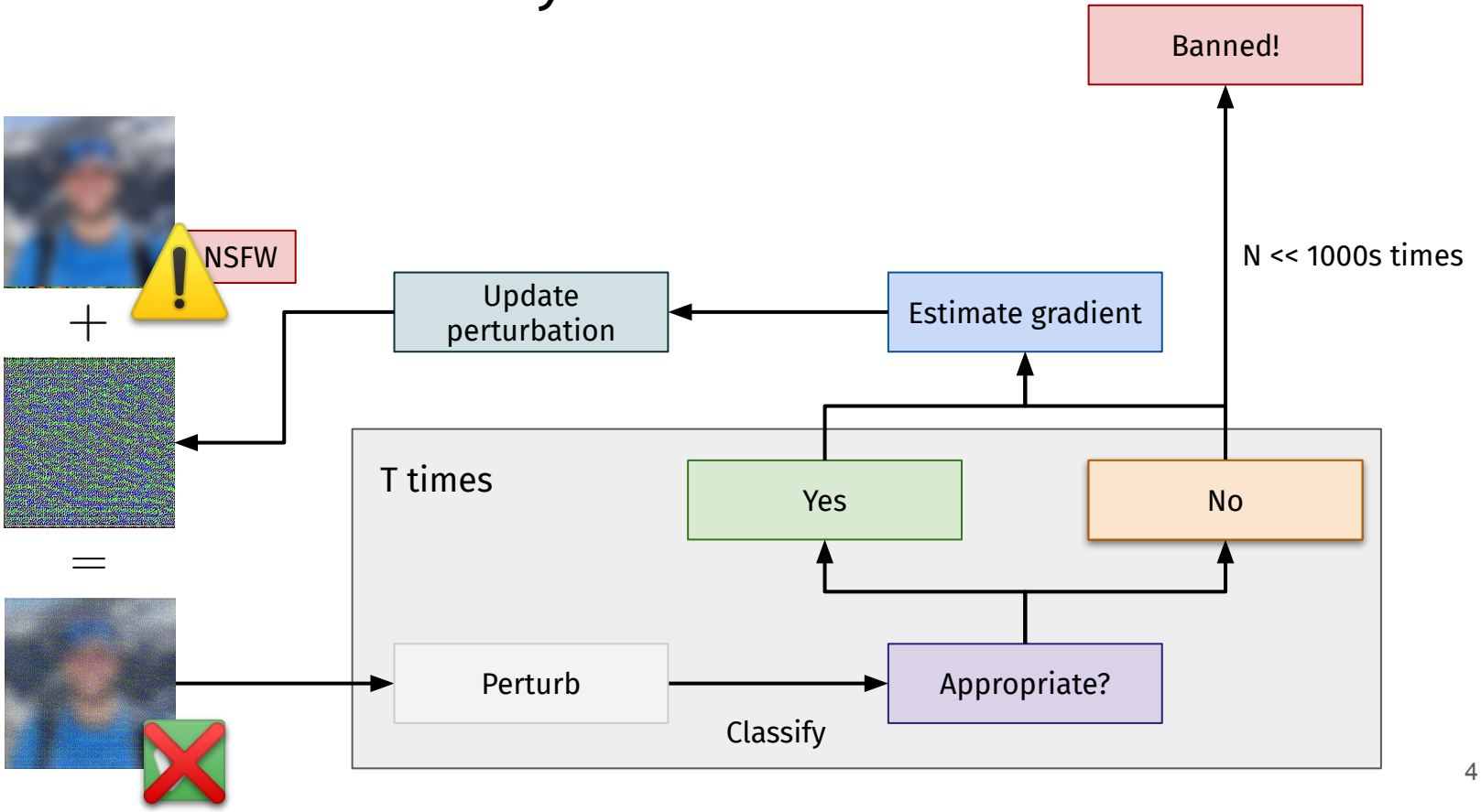
We can create an adversarial example, i.e., craft a small perturbation that fools the model!



How can we upload an inappropriate image on Twitter?

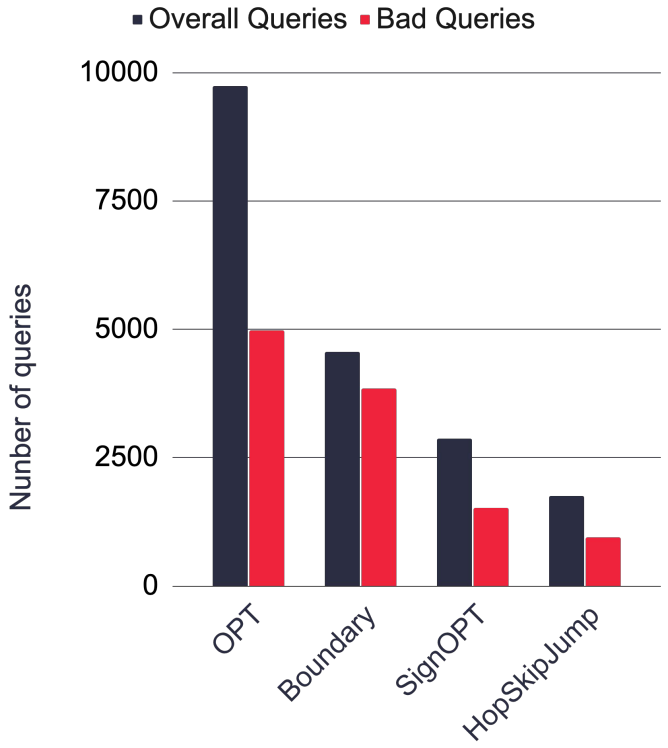


How do current label-only attacks work?

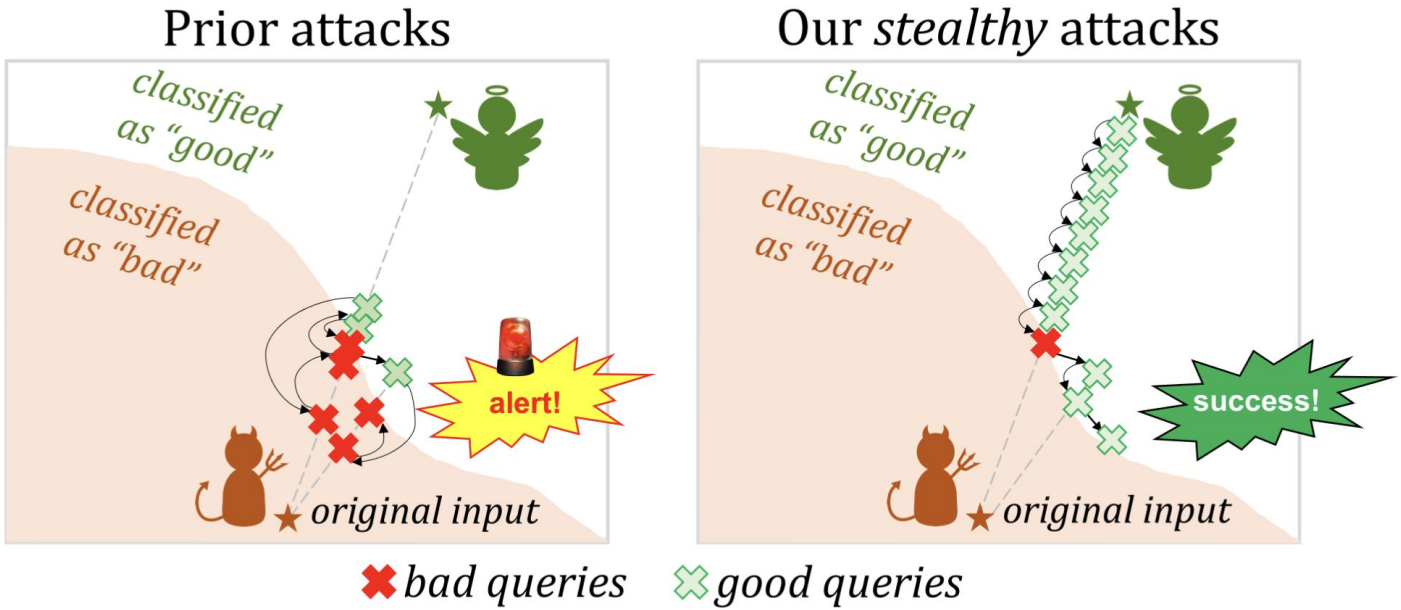


Can we use existing evasion attacks to upload an inappropriate image on Twitter? **No.**

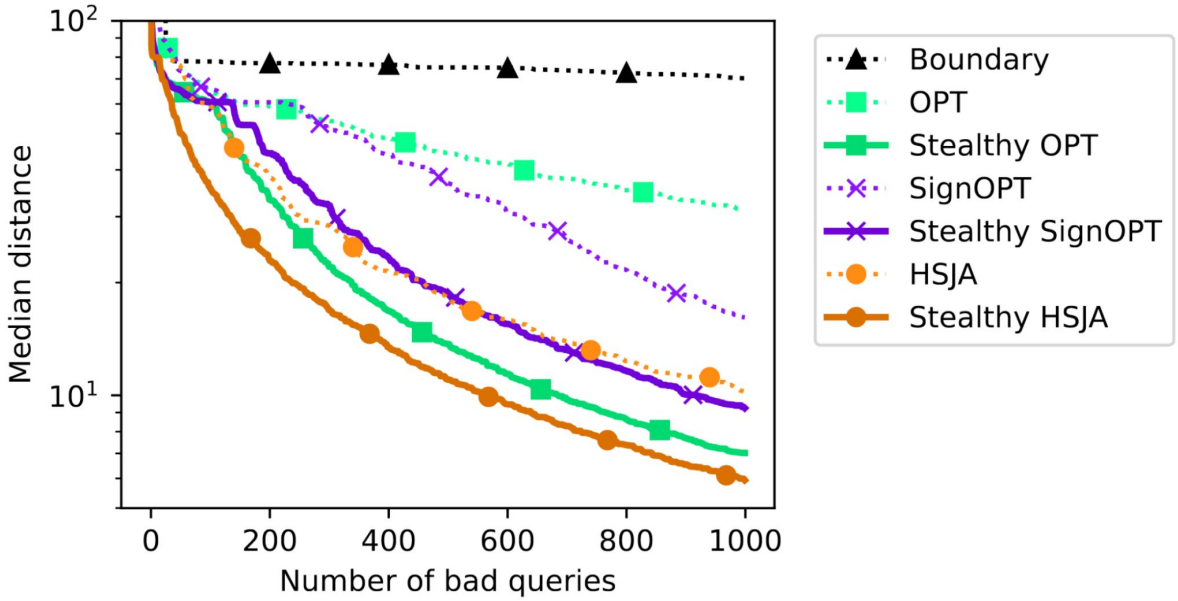
- Existing attacks do a lot (100s) of queries that are classified as “bad”.
- Any user would get **banned** after very few such queries
- There is an **asymmetric cost** between “good” and “bad” queries



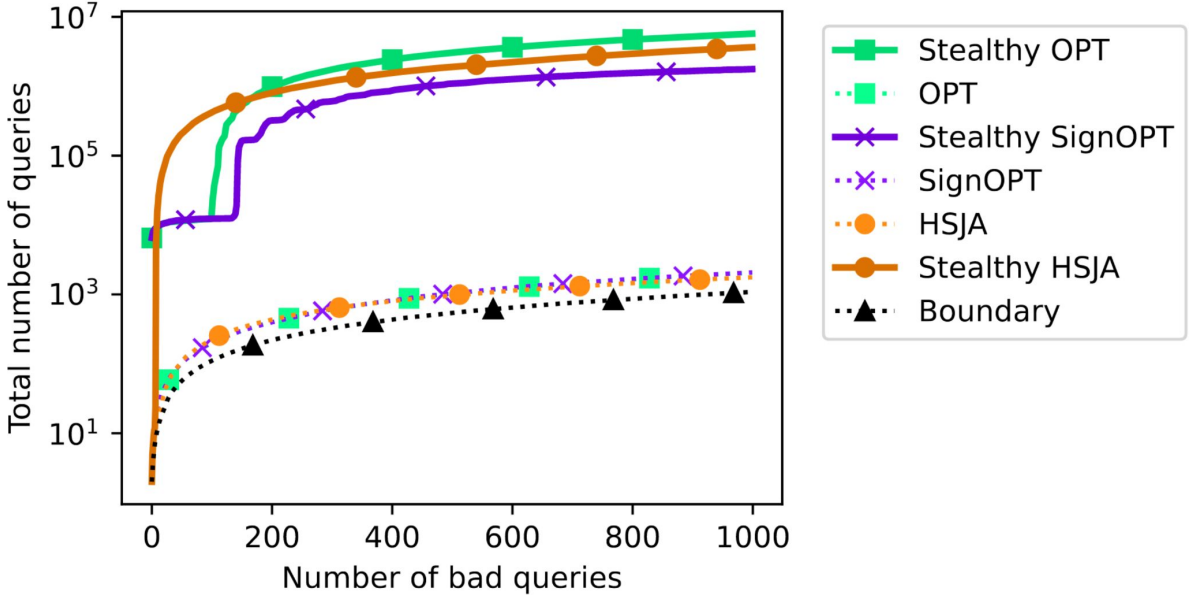
How do we make attacks stealthier?



Is this enough?



Is this enough? **No.**



Takeaways

- Only counting the overall number of queries is not enough for many real-world applications
- There is an asymmetric cost between “good” and “bad” queries
- It’s possible to stealthy-fy existing attacks, but it comes at a cost in terms of overall queries

We should rethink label-only attacks from scratch

edebenedetti@inf.ethz.ch
@edoardo_debe

Paper

Leaderboard