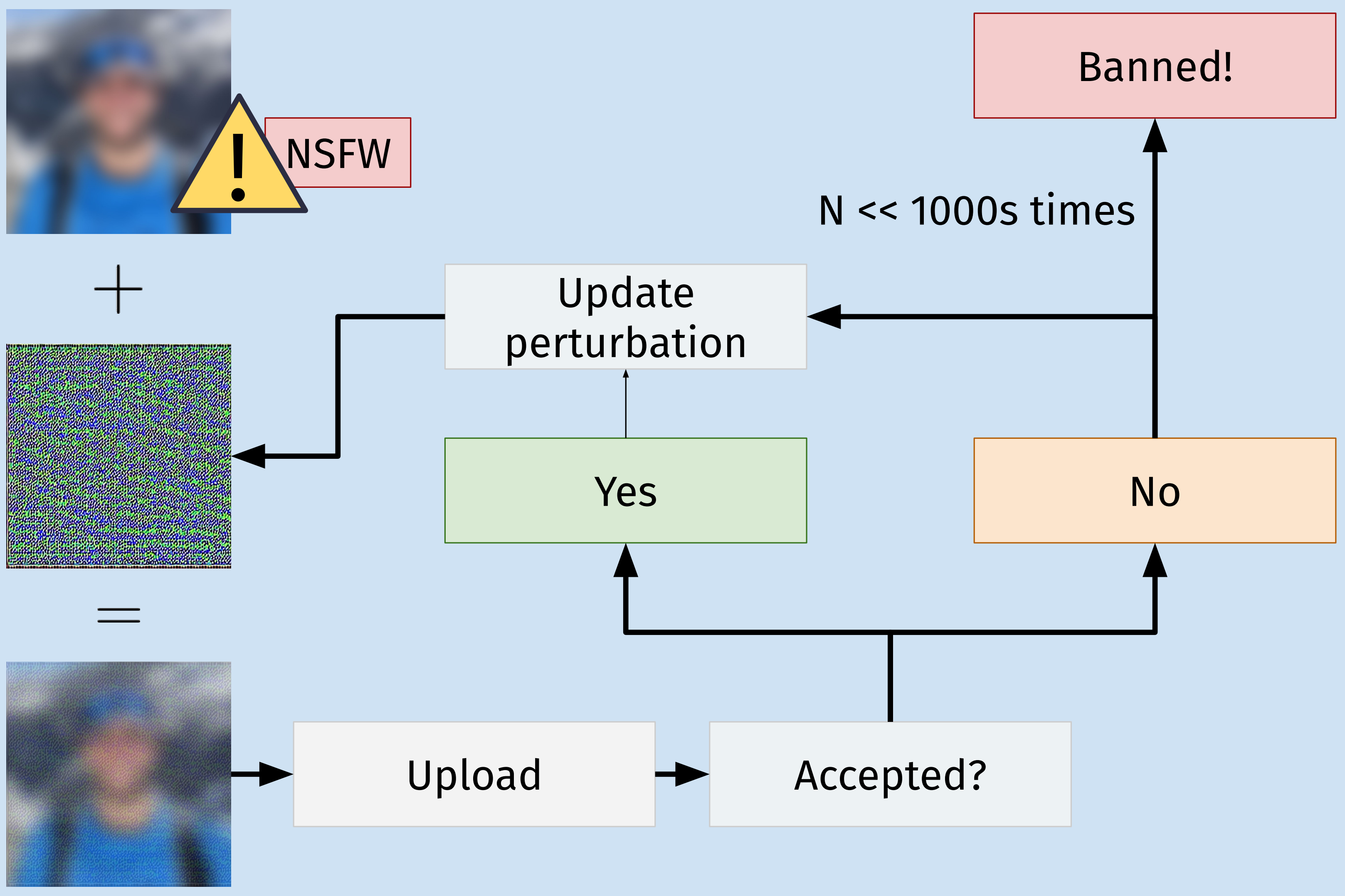
ETH zürich | Google DeepMind

Evading Black-box Classifiers Without Breaking Eggs

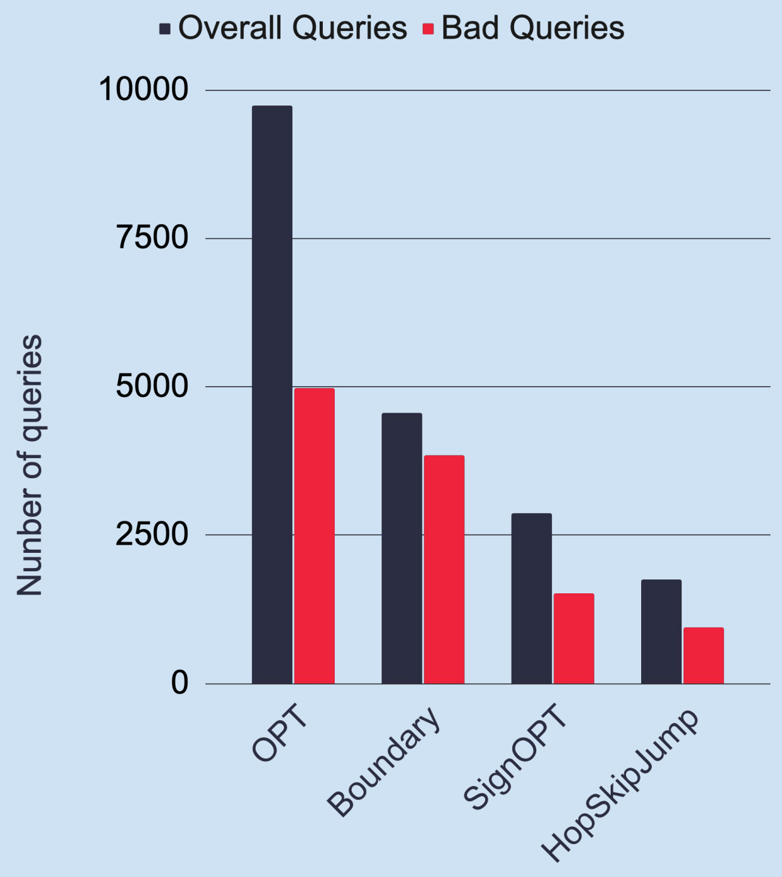
Paper   Code

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

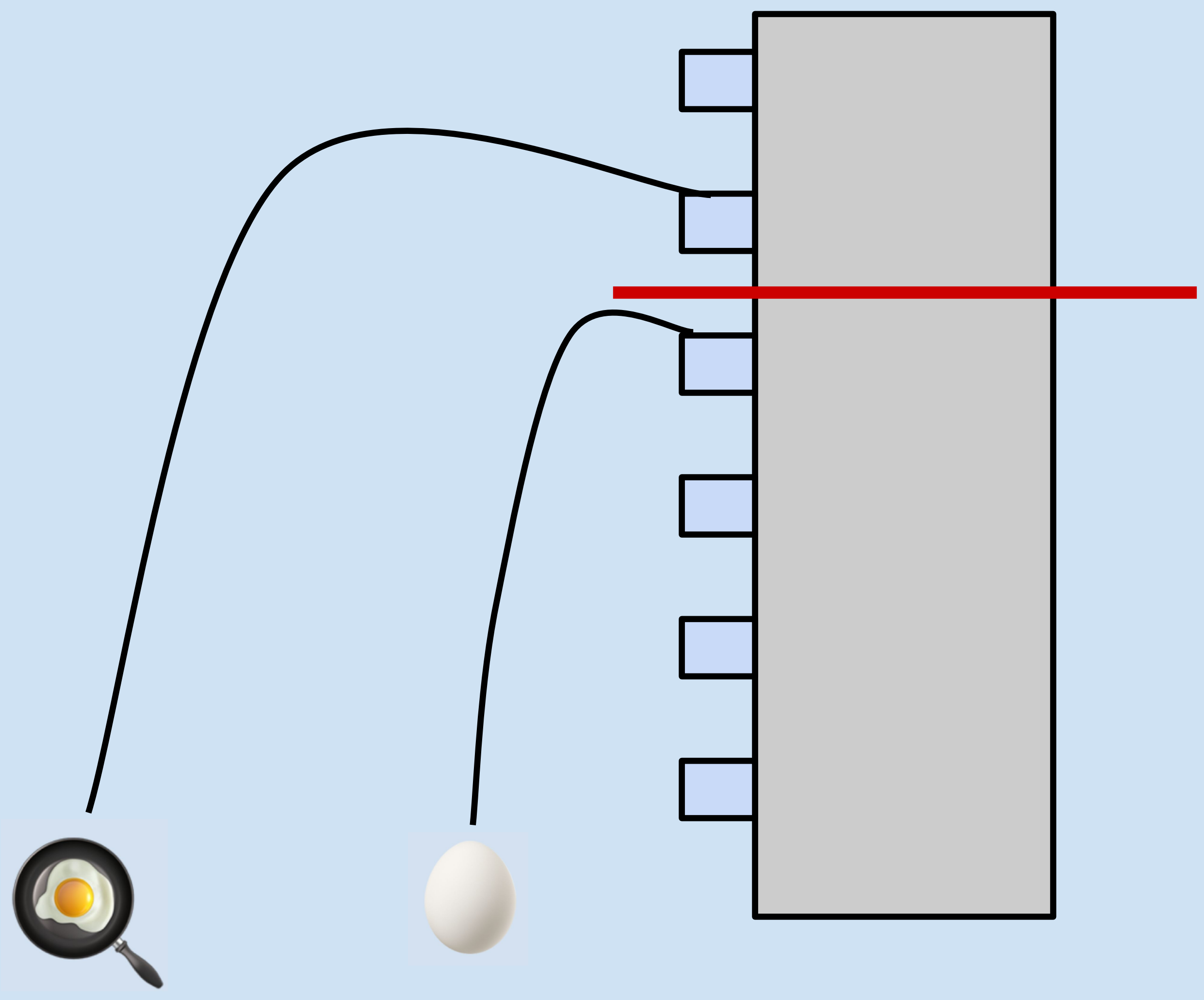
1/ Label-only adversarial attacks are effective, but make many “bad” queries



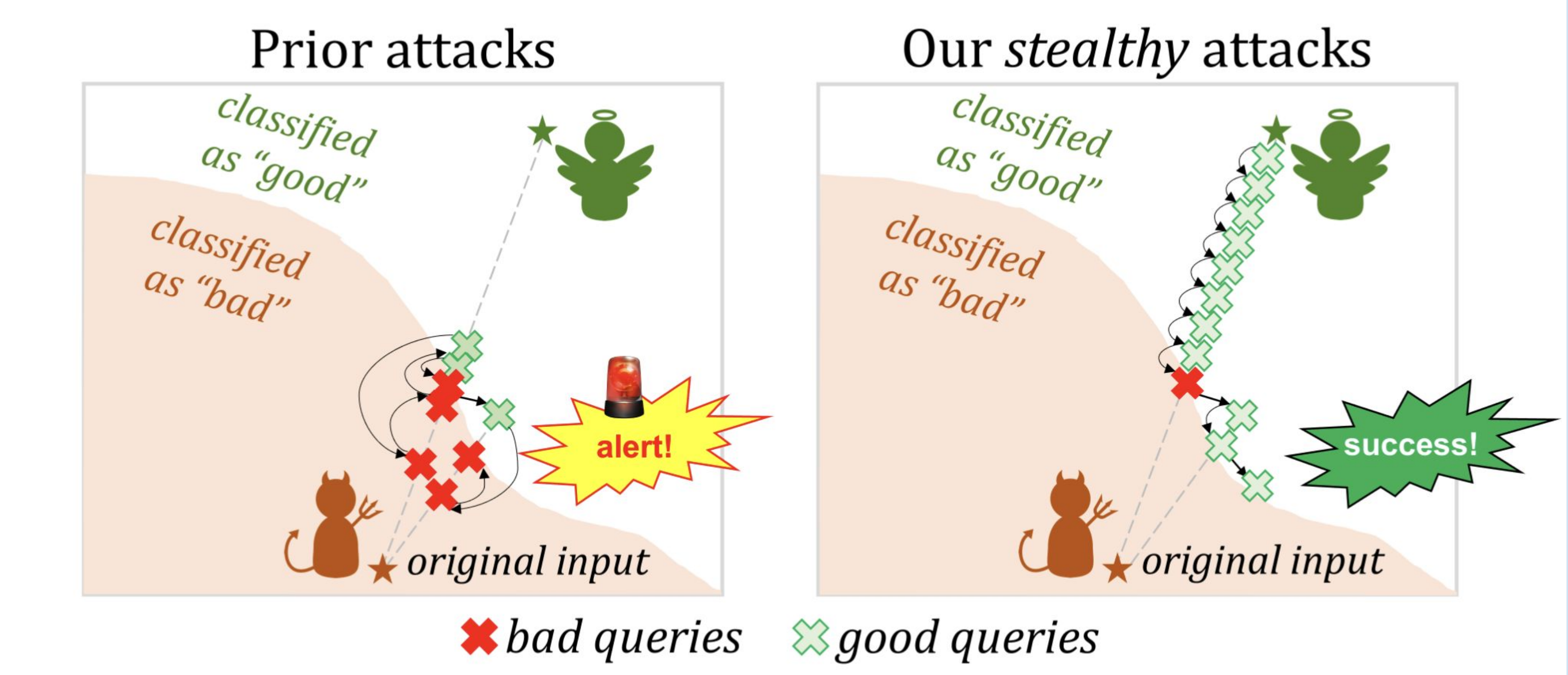
- Current attacks optimize for the total number of queries
- This makes them unusable for many real-world applications, as an adversary would be banned after very few “bad” queries
- There is a clear asymmetric cost between “good” and “bad” queries



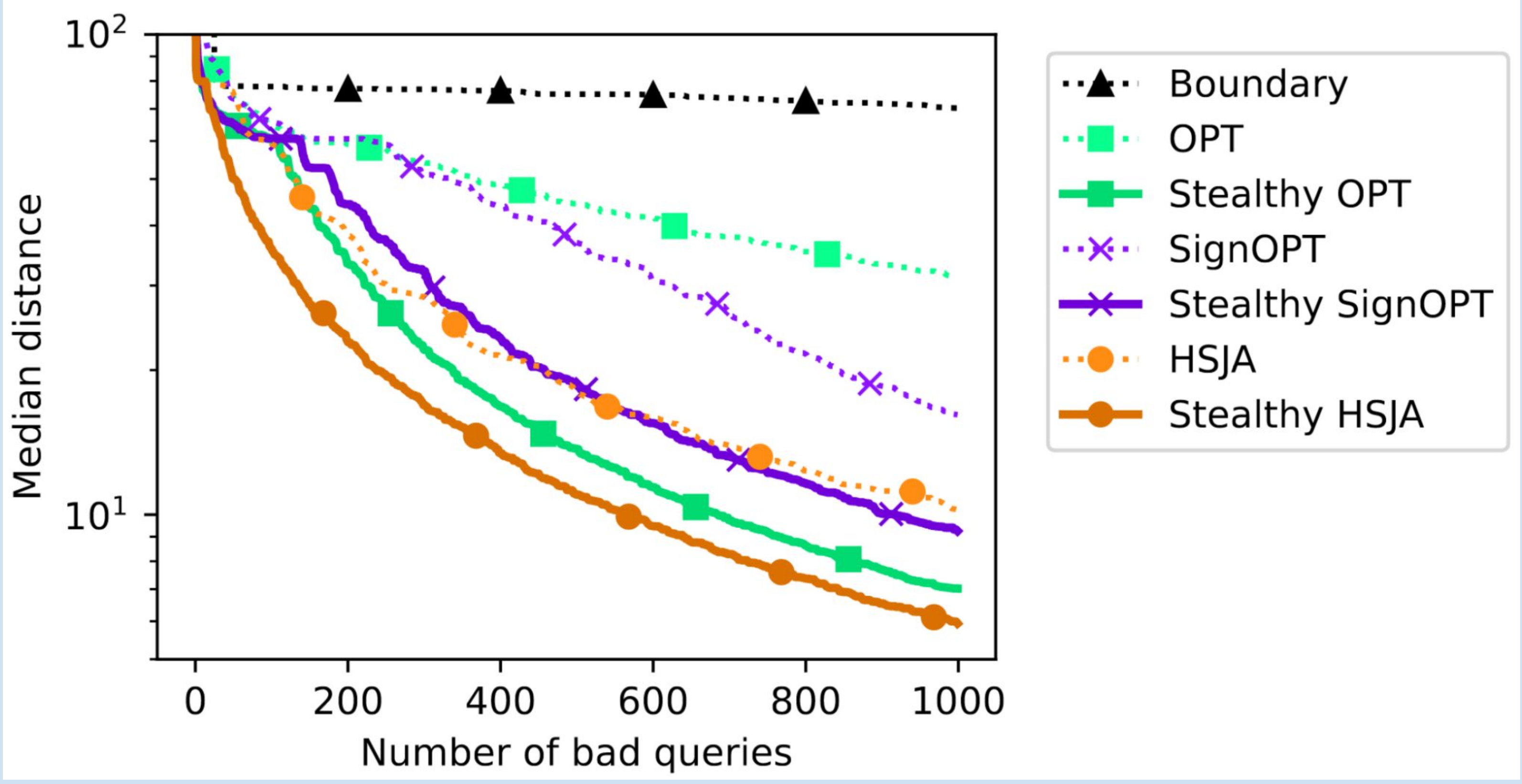
2/ A “bad” query is like breaking an egg



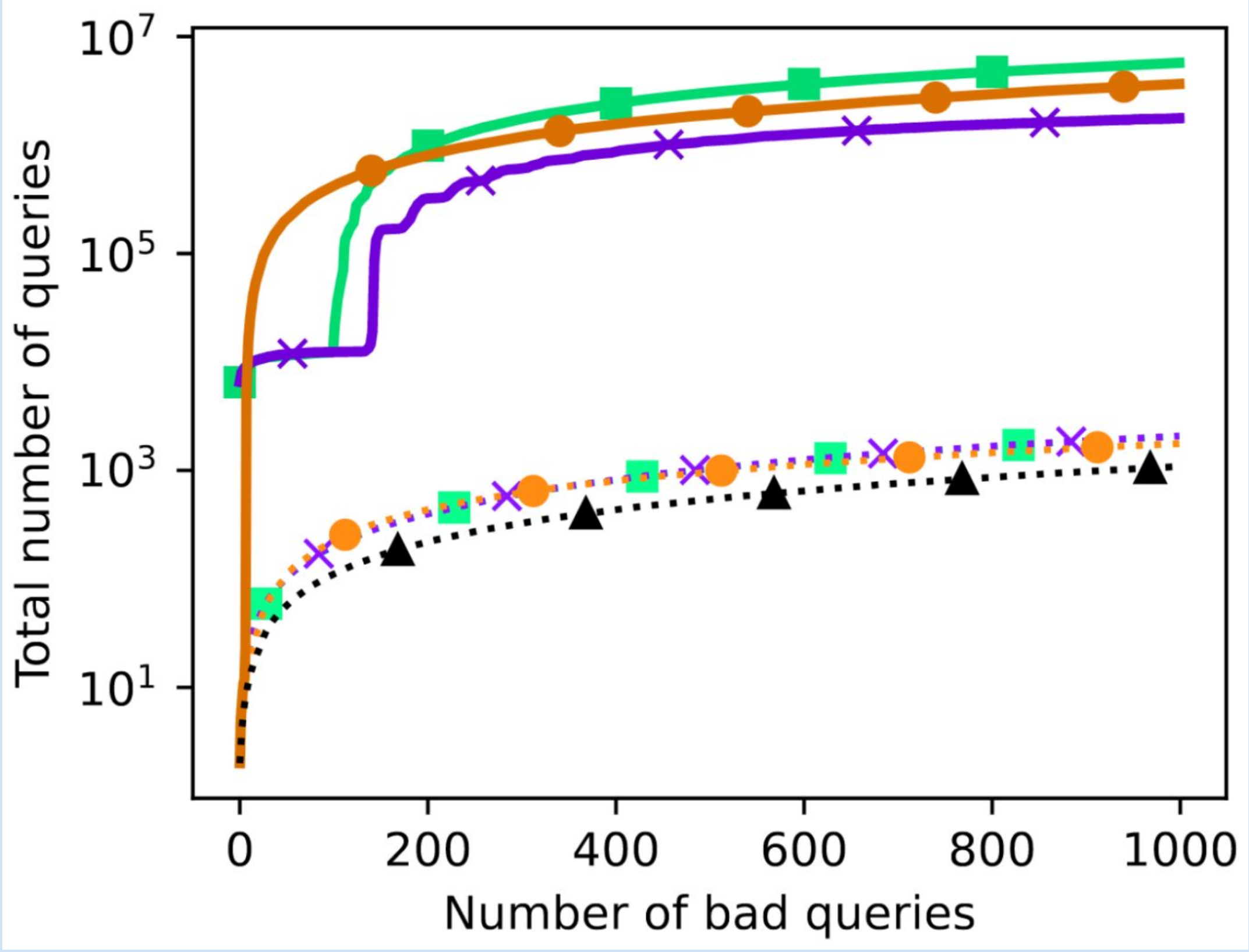
3/ How do we make our attacks “stealthy”?



4/ Our attacks make fewer “bad” queries



5/ … but many more total queries



Can you do better?