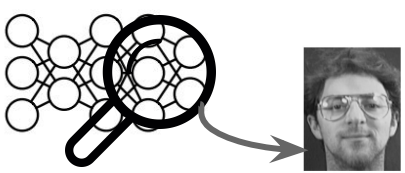
Types of adversaries and our threat model

Model querying (**black-box adversary**)

Shokri et al. (2016) *Membership Inference Attacks against ML Models*
Fredrikson et al. (2015) *Model Inversion Attacks*



Model inspection (**white-box adversary**)

Zhang et al. (2017) *Understanding DL requires rethinking generalization*

In our work, the threat model assumes:

- Adversary can make a potentially unbounded number of queries
- Adversary has access to model internals