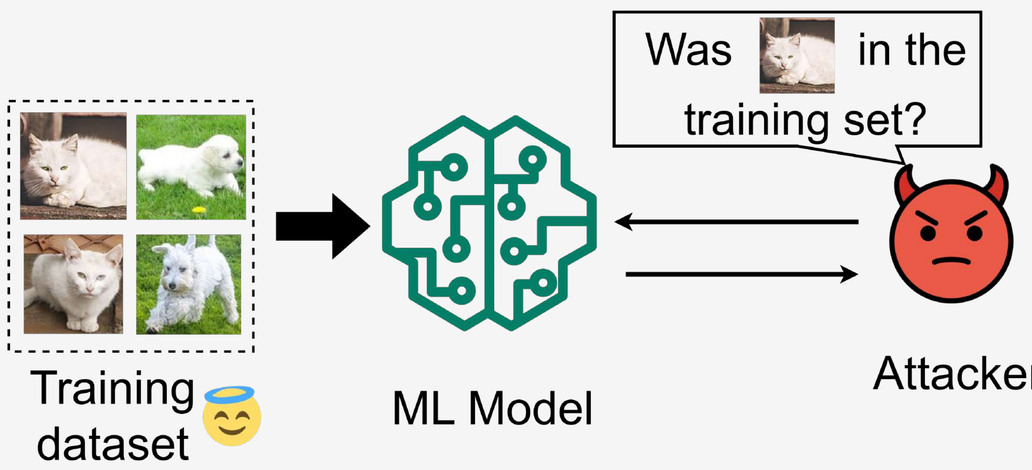
Evaluating Membership Inference Attacks in heterogeneous-data setups

Bram van Dartel, Marc Damie, Florian Hahn

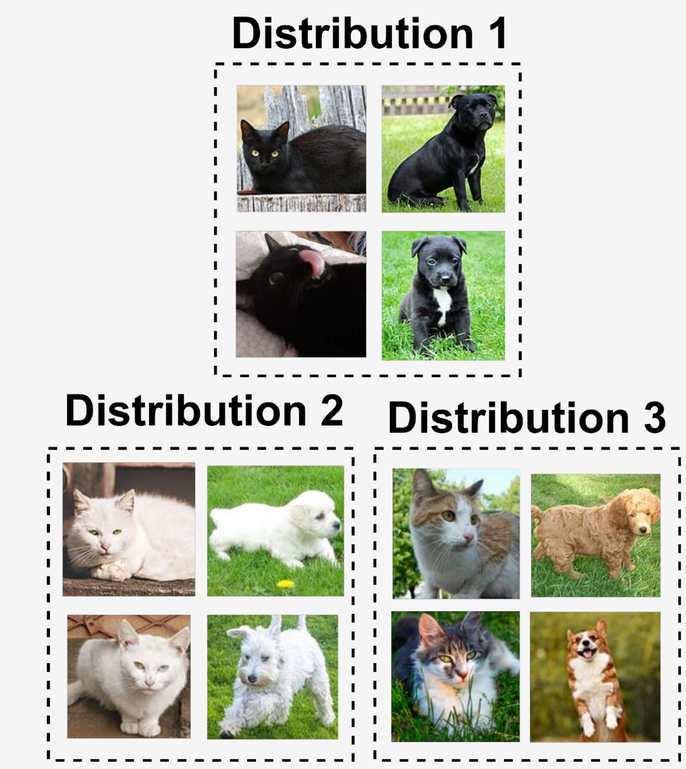
Membership Inference Attack (MIA)



Attacker's knowledge

Target ML model + A "**similar**" **dataset**

Data heterogeneity in ML



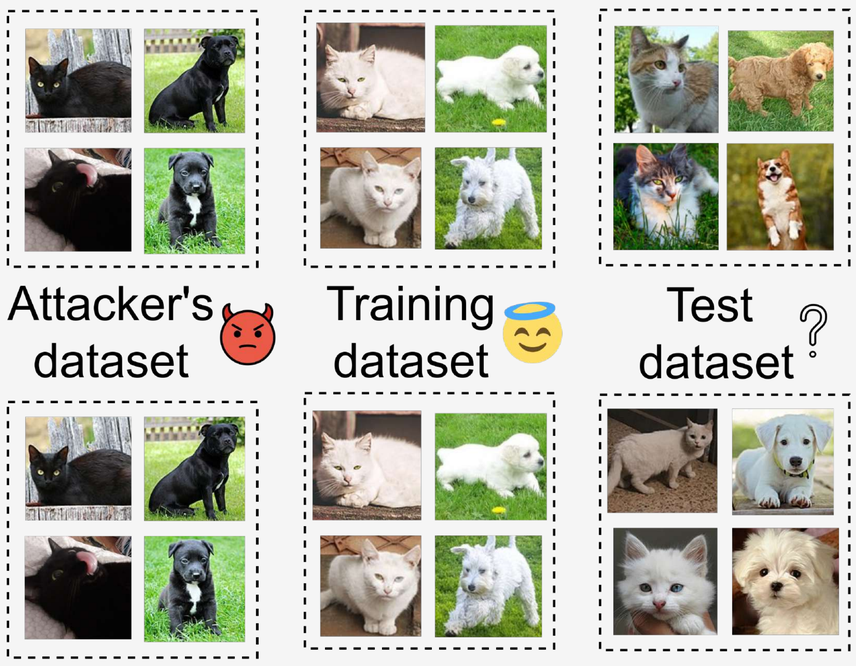
Research question

How accurate are Membership Inference Attacks in heterogeneous-data setups?

How to simulate data heterogeneity in MIAs?

Option 1:
Three-distribution setting
(used in existing works)

Option 2:
Two-distribution setting



Results on tabular data

➡ **91%**

Attack Accuracy

➡ **47%**

Takeaways

- Option 1 ≈ "**distribution membership** inference attack."
- A given attack can be totally **efficient or inefficient depending on the simulation settings**...
- Open problem: What is a **realistic attack setup** for MIA?

Funding

Supported by NWO under **NWO:SHARE project [CS.011]**.

UNIVERSITY OF TWENTE.