DAWN: Dynamic Adversarial Watermarking of DNNs

Can we verify model ownership after model extraction?

The setting:

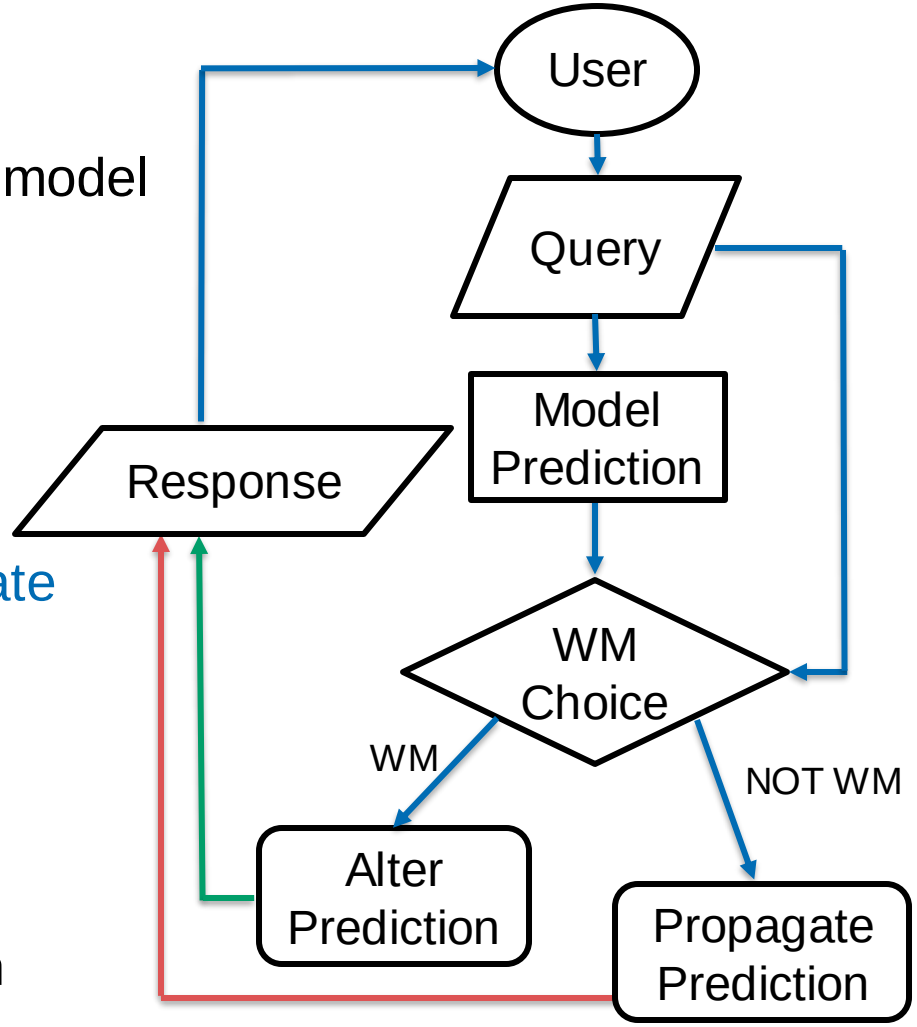
- Malicious client can use query responses to train surrogate model
- Machine learning models constitute business advantage
- Prevention and detection not reliable

Our approach - deter using watermarking:

- Return incorrect predictions for selected inputs
- Force adversary to embed watermark while training surrogate
- Demonstrate ownership when model is exposed

Properties and challenges:

- Desiderata: unremovable, indistinguishable, reliable ownership demonstration
- Challenges: double extraction, robust "WM Choice" function



In submission