2023 - AI - **252-066** - DisGUIDE: Disagreement-Guided Data-Free Model Extraction - rosenth0@purdue.edu - Jonathan Rosenthal

CERIAS

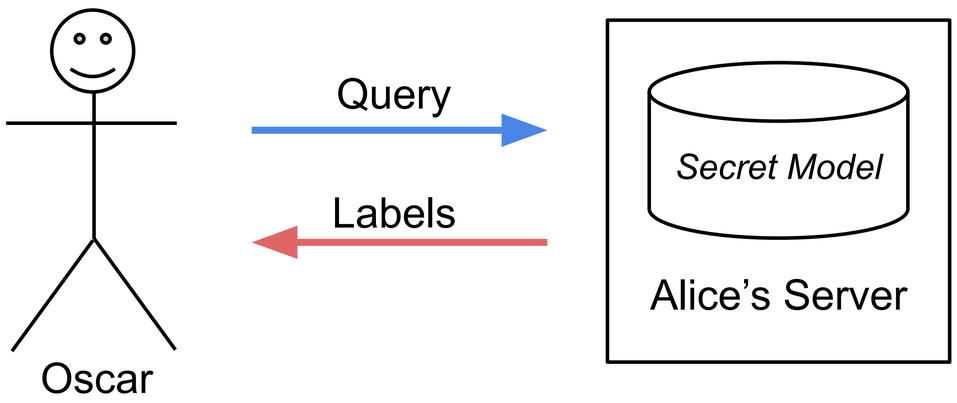
The Center for Education and Research in Information Assurance and Security

DisGUIDE: Disagreement-Guided Data-Free Model Extraction

Jonathan Rosenthal[1], Eric Enouen[2], Hung Viet Pham[3], Lin Tan[1]

[1] Purdue University [2] The Ohio State University [3] York University

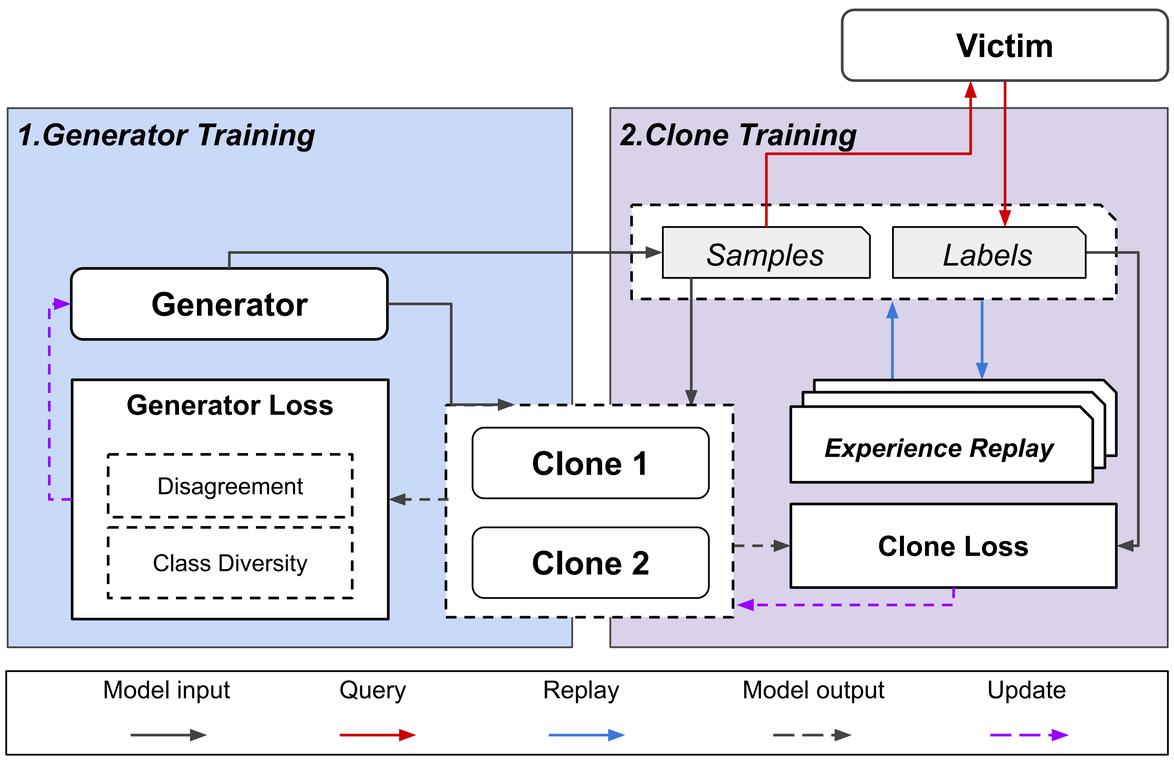
Can Oscar copy Alice's secret model?



Oscar has no data and can only query the server in a blackbox fashion!

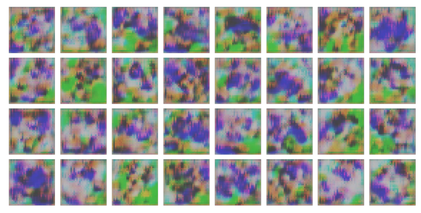
How did we approach this problem?

Model	Prediction
Clone A	Dog
Clone B	Cat



- Novel generator loss: **Clone disagreement**
  - No victim access required!
- General method with minimal constraints!
- Experience replay augments clone training.

Generated images do not resemble victim training data!



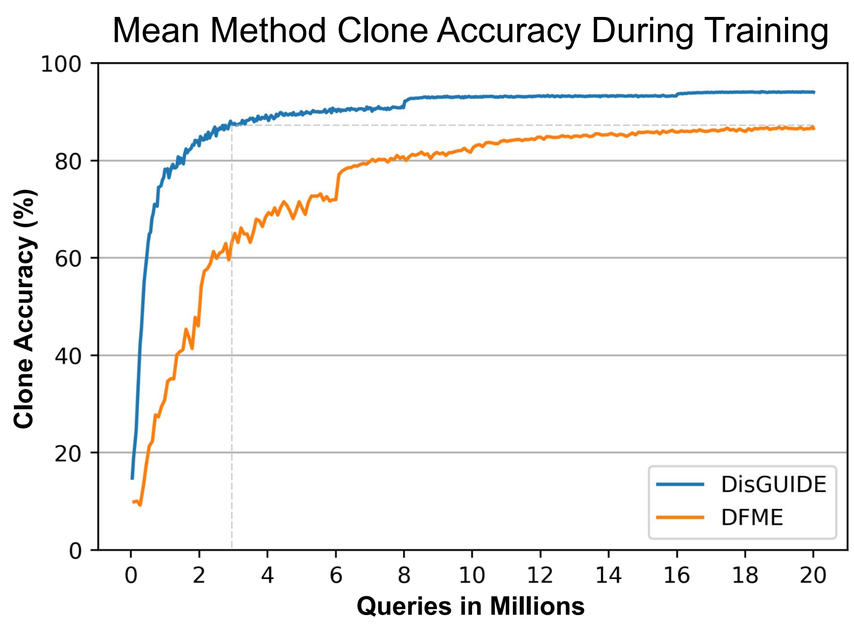
How well did we solve this problem?

DisGUIDE improves final model accuracy!

		CIFAR-10, Victim ~95.5%		CIFAR-100, Victim ~78%	
Setting	Technique	Budget	Accuracy (%)	Budget	Accuracy (%)
Soft-Label	DFME	20M	88.10	/	/
	DFMS	20M	91.24	/	/
	DisGUIDE	20M	**94.02** ± 0.25	10M	**69.47** ± 0.88
Hard-Label	DFMS	8M	84.51	10M	43.56
	DisGUIDE	8M	**87.93** ± 1..74	10M	**62.04** ± 1.03

First CIFAR-100 SL results!

95% Confidence Interval



DisGUIDE improves query efficiency!

65% fewer queries to reach prior SotA for CIFAR-10!

Conclusions and Future Work

- DisGUIDE improves SotA for model extraction accuracy and query efficiency!
- DisGUIDE leverages clone disagreement for improved model extraction!
- Data-free model-extraction is a challenging task with much room for improvement!

Paper QR Codes

Paper

Code

PURDUE UNIVERSITY

CERIAS