U.S. NAVAL RESEARCH LABORATORY

Distinguished Paper Award, Internet Defense 2nd Prize!

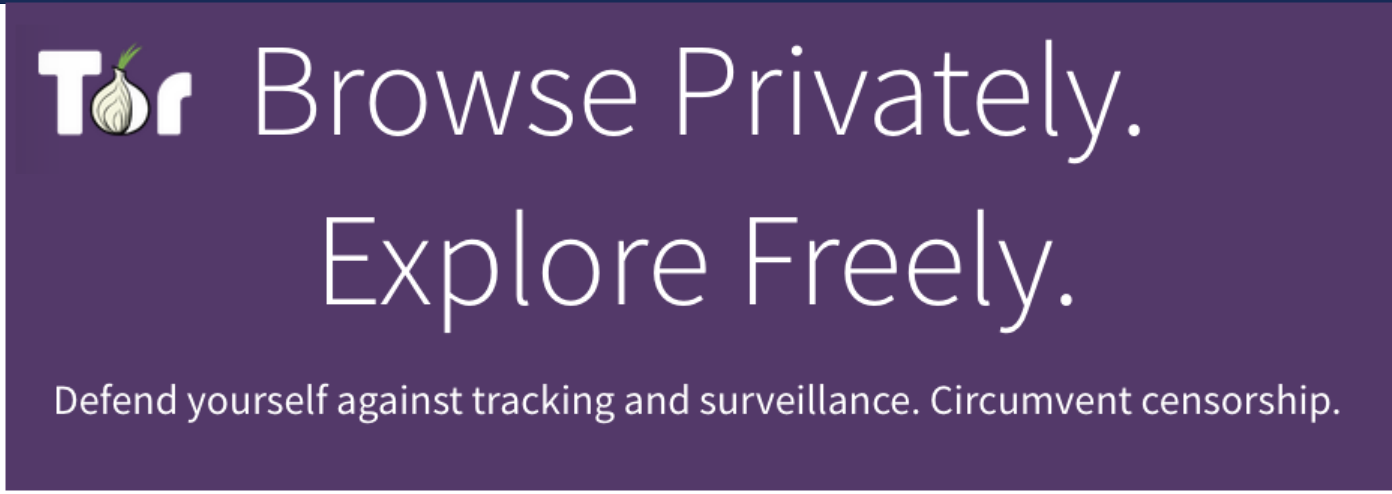
Online Website Fingerprinting: Evaluating Website Fingerprinting Attacks on Tor in the Real World

Giovanni Cherubin, ~~Alan Turing Institute~~ Microsoft Research
Rob Jansen, U.S. Naval Research Laboratory
Carmela Troncoso, EPFL SPRING Lab

Rob Jansen, Ph.D.
Computer Security Research Scientist
Center for High Assurance Computer Systems
U.S. Naval Research Laboratory

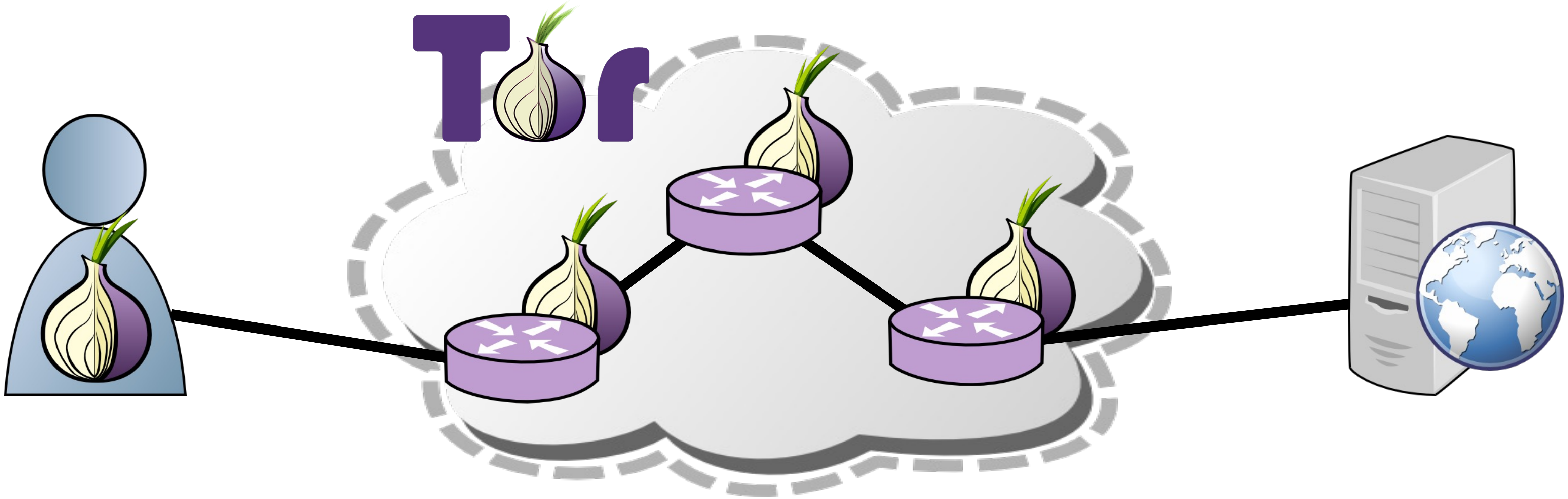
31st USENIX Security Symposium
Boston, MA, USA
August 10th, 2022

How Tor Works



Anonymous Communication and Tor

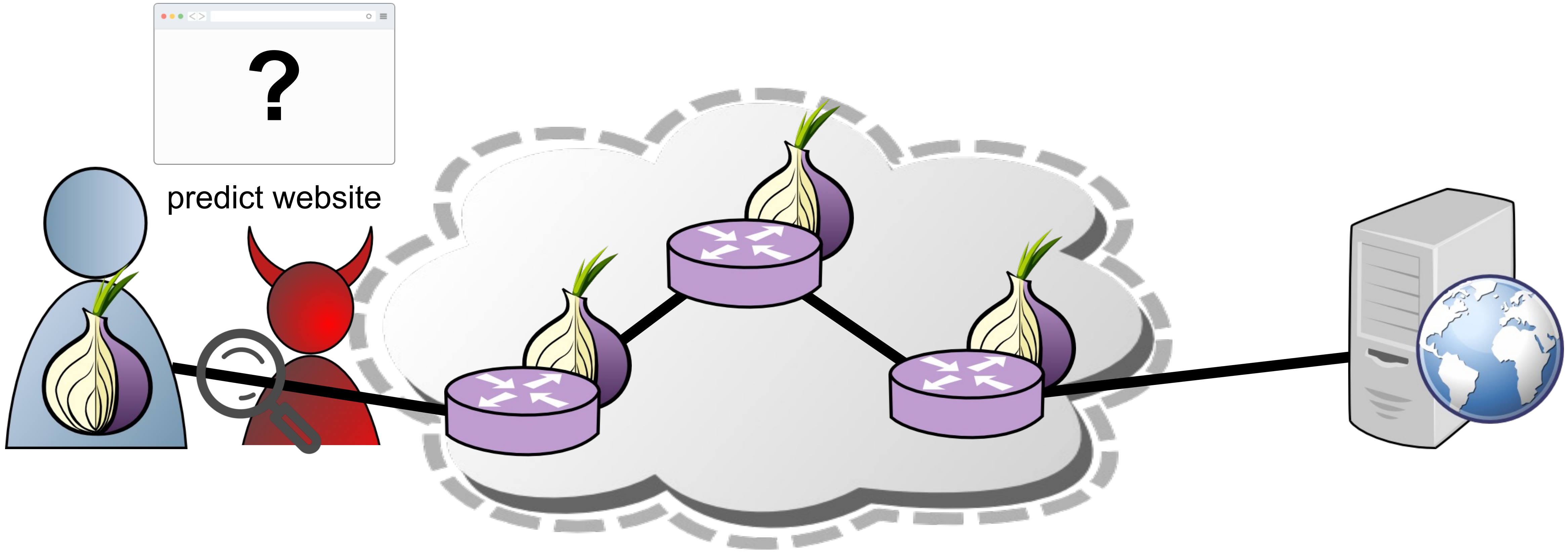
- Separates identification from routing
- Provides unlinkable communication
- Promotes user safety and privacy online



Deanonymizing Tor Users

Website fingerprinting attack

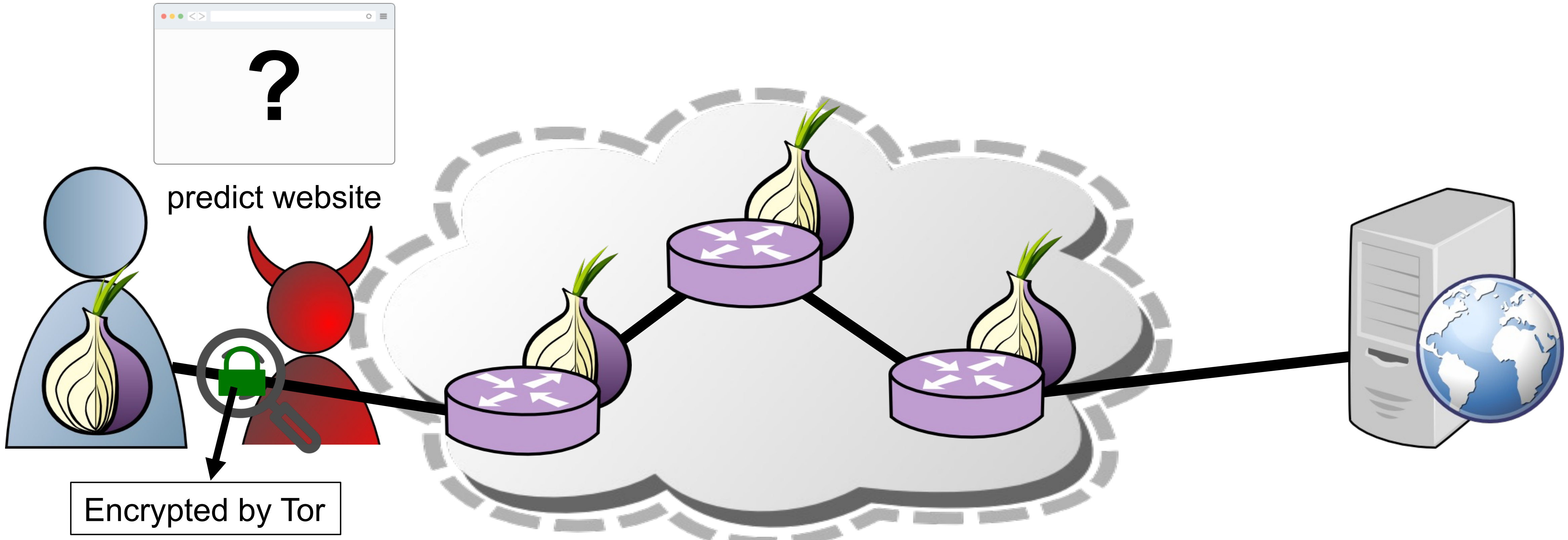
- Predict website visited by user
- Requires access to <u>entry side only</u>



Deanonymizing Tor Users

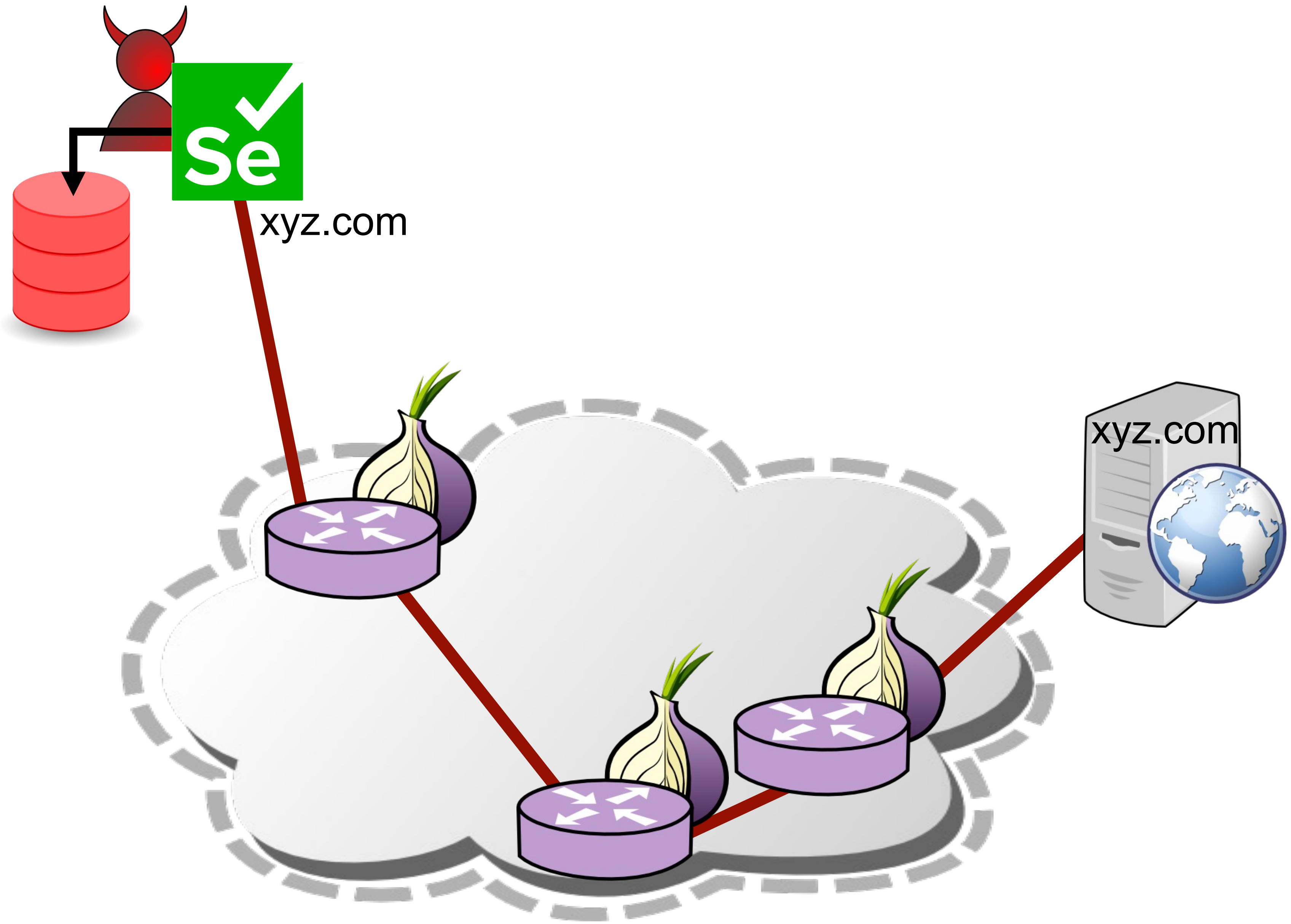
Website fingerprinting attack
- Predict website visited by user
- Requires access to entry side only

Problem:
- Need labels to train ML classifiers for website prediction
- Genuine labels are encrypted



Website Fingerprinting Threat Model

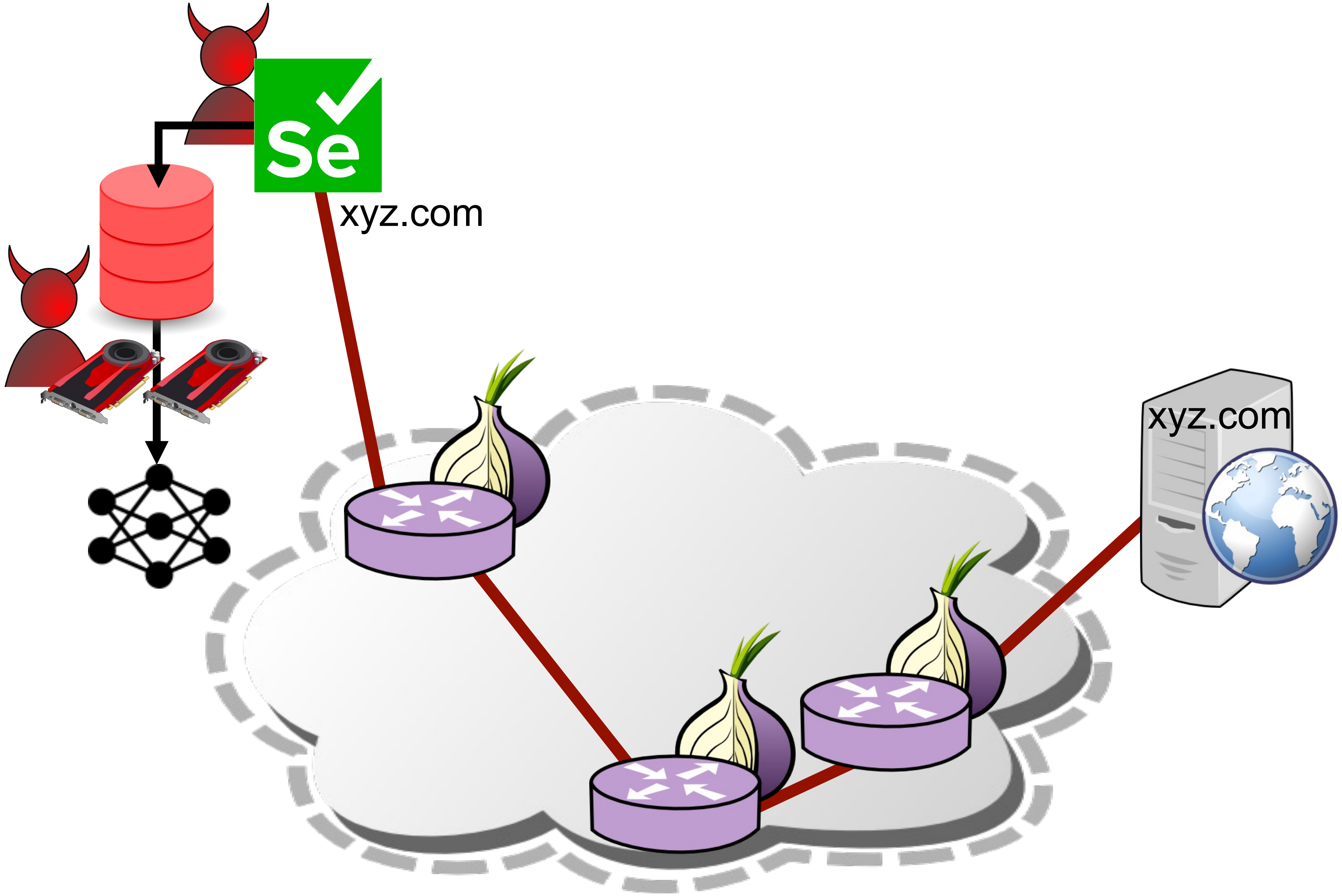
Step 1: gather data & labels
- Use automated browser (selenium) to crawl websites



Website Fingerprinting Threat Model

Step 1: gather data & labels
- Use automated browser (selenium) to crawl websites

Step 2: train ML classifier
- Use collected data & labels

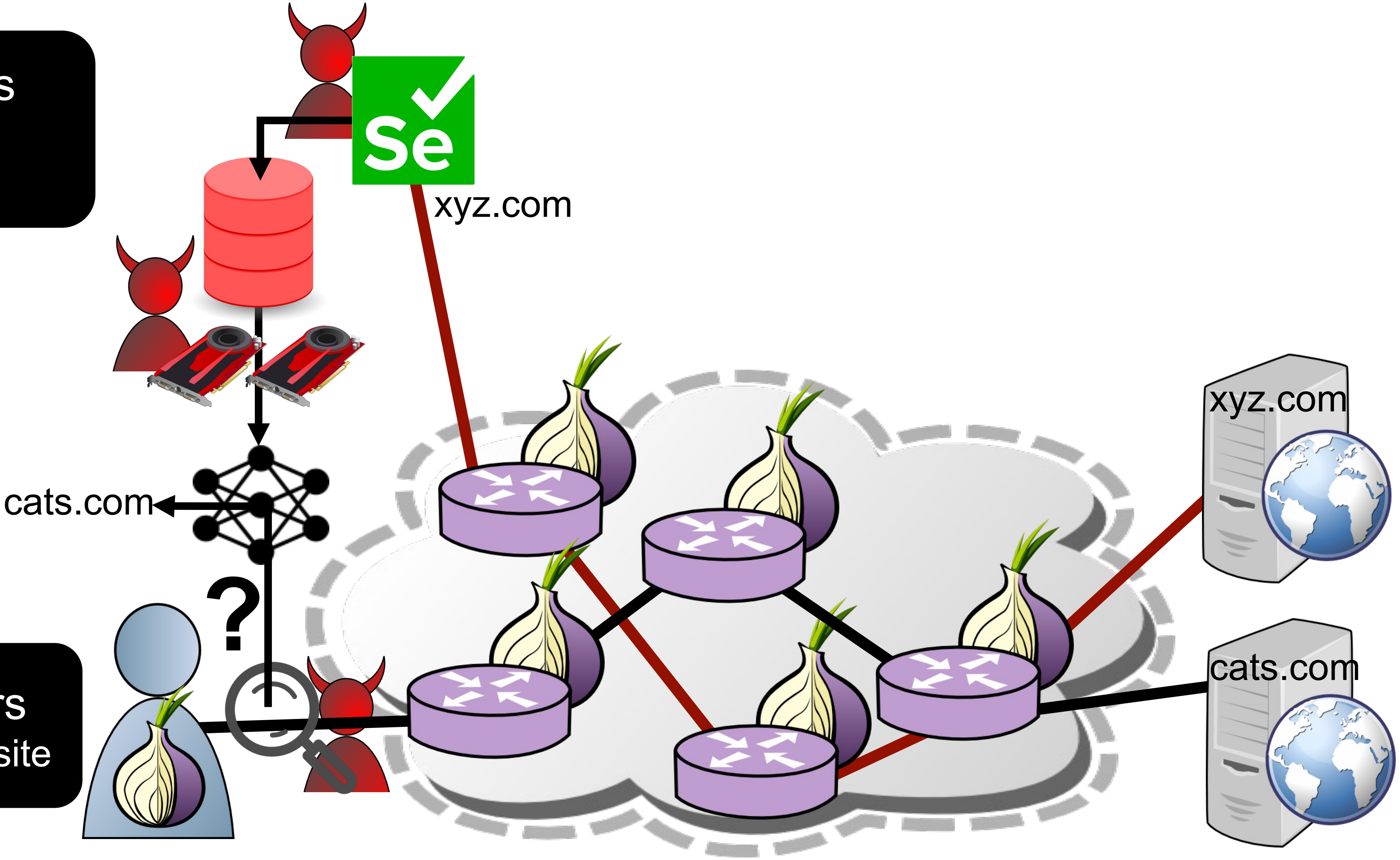


Website Fingerprinting Threat Model

Step 1: gather data & labels
- Use automated browser (selenium) to crawl websites

Step 2: train ML classifier
- Use collected data & labels

Step 3: deploy against users
- Use ML model to predict website

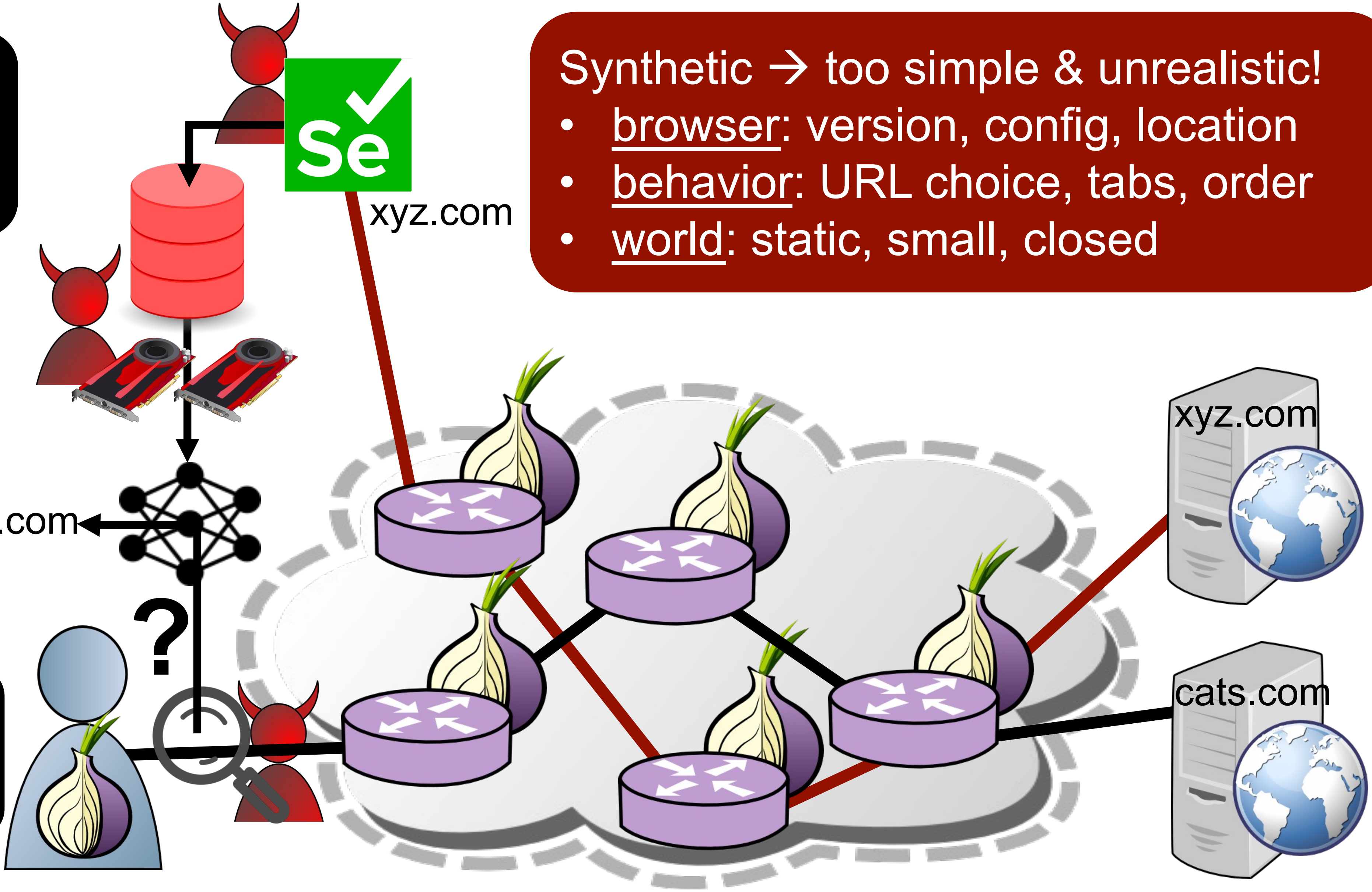


Criticisms of Website Fingerprinting Threat Model

Step 1: gather data & labels
- Use automated browser (selenium) to crawl websites

Step 2: train ML classifier
- Use collected data & labels

Step 3: deploy against users
- Use ML model to predict website



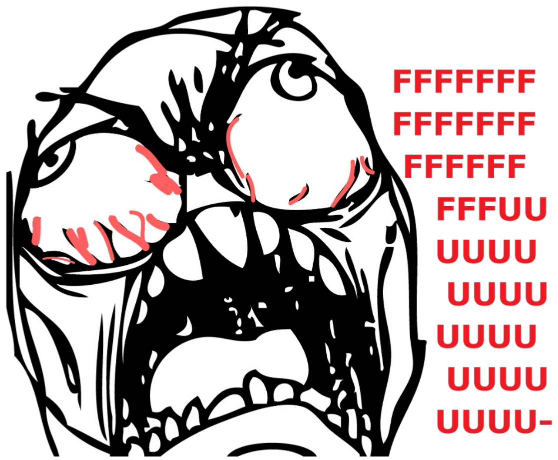
Synthetic → too simple & unrealistic!
- browser: version, config, location
- behavior: URL choice, tabs, order
- world: static, small, closed

What is the threat of WF attacks in the *real* world?

~~Synthetic model~~

- ~~Overly simple and unrealistic~~
- ~~High ML accuracy in simple model~~

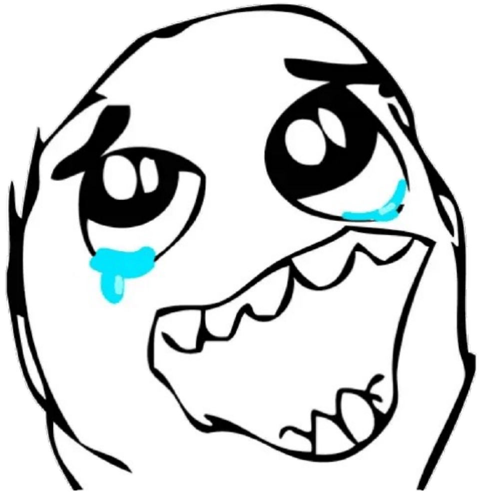
Stop using!!



Genuine model

- Consider genuine data & labels from a Tor exit relay

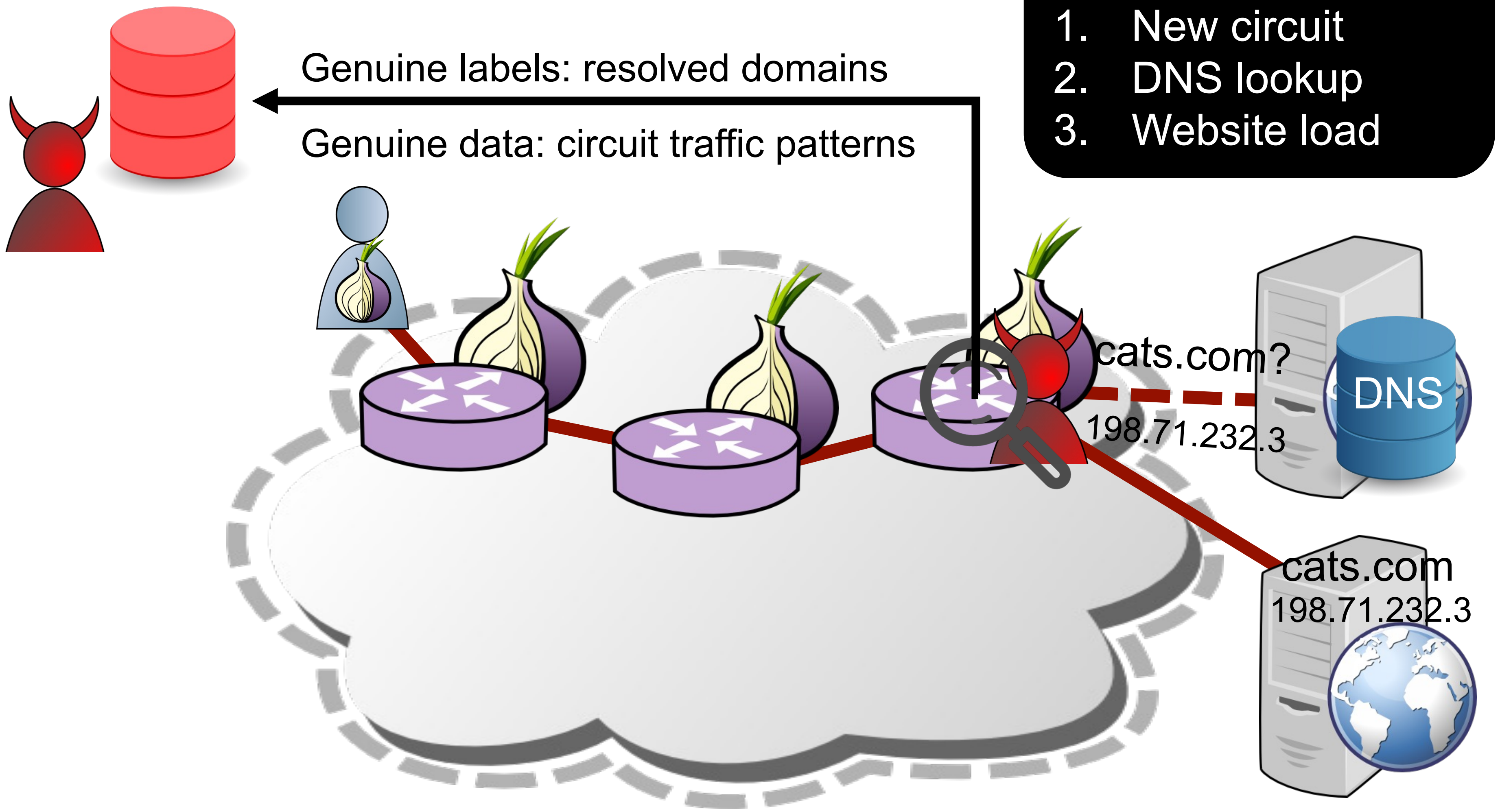
Our new approach



Key Insight: Exits Observe Genuine Data & Labels

Step 1: gather data & labels

- Run a Tor exit relay and use to to collect genuine Tor traffic



Key Insight: Exits Observe Genuine Data & Labels

Step 1: gather data & labels

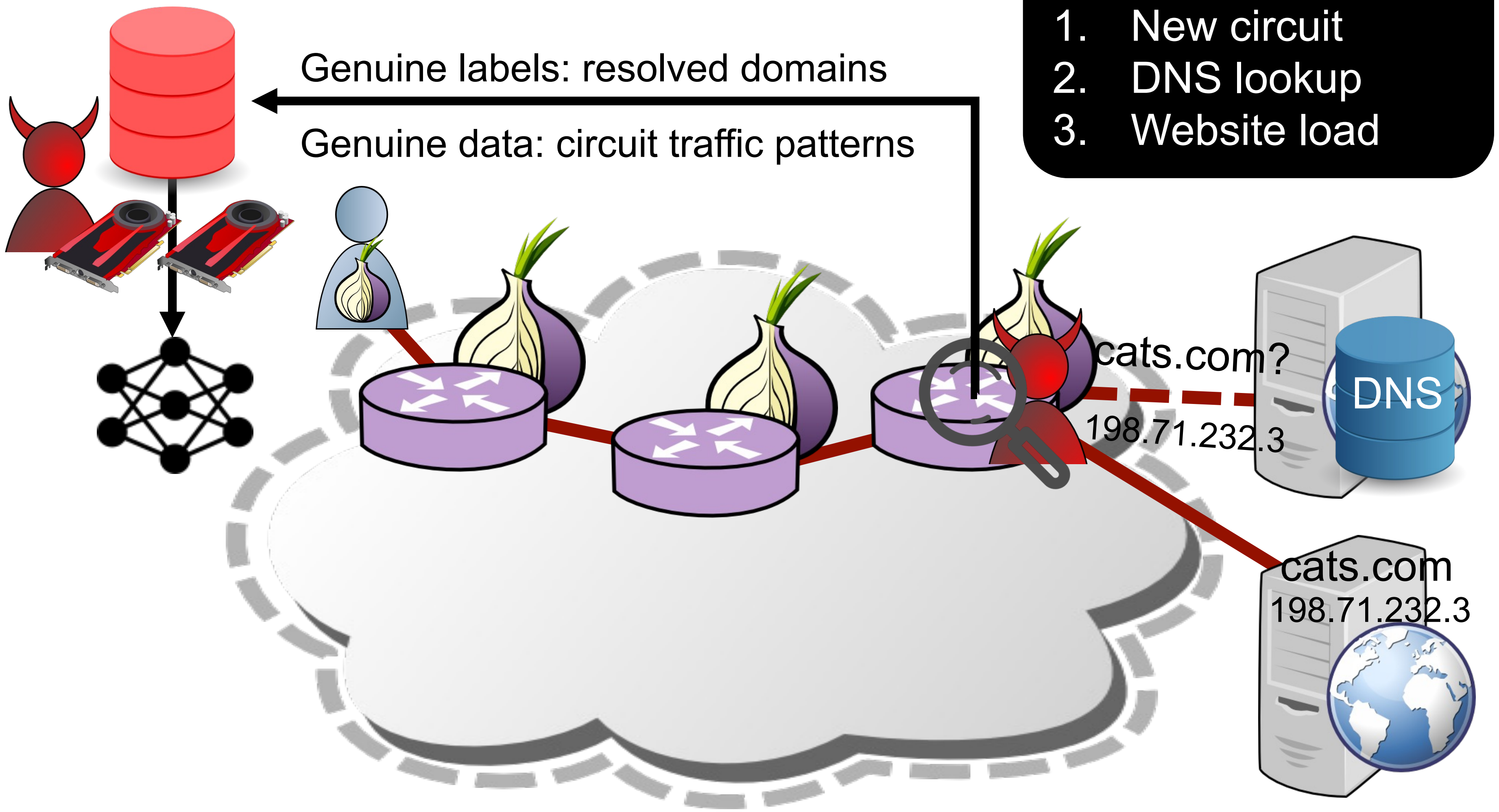
- Run a Tor exit relay and use to to collect genuine Tor traffic

Step 2: train ML classifier

- Use collected data & labels

Exit can observe:

1. New circuit
2. DNS lookup
3. Website load



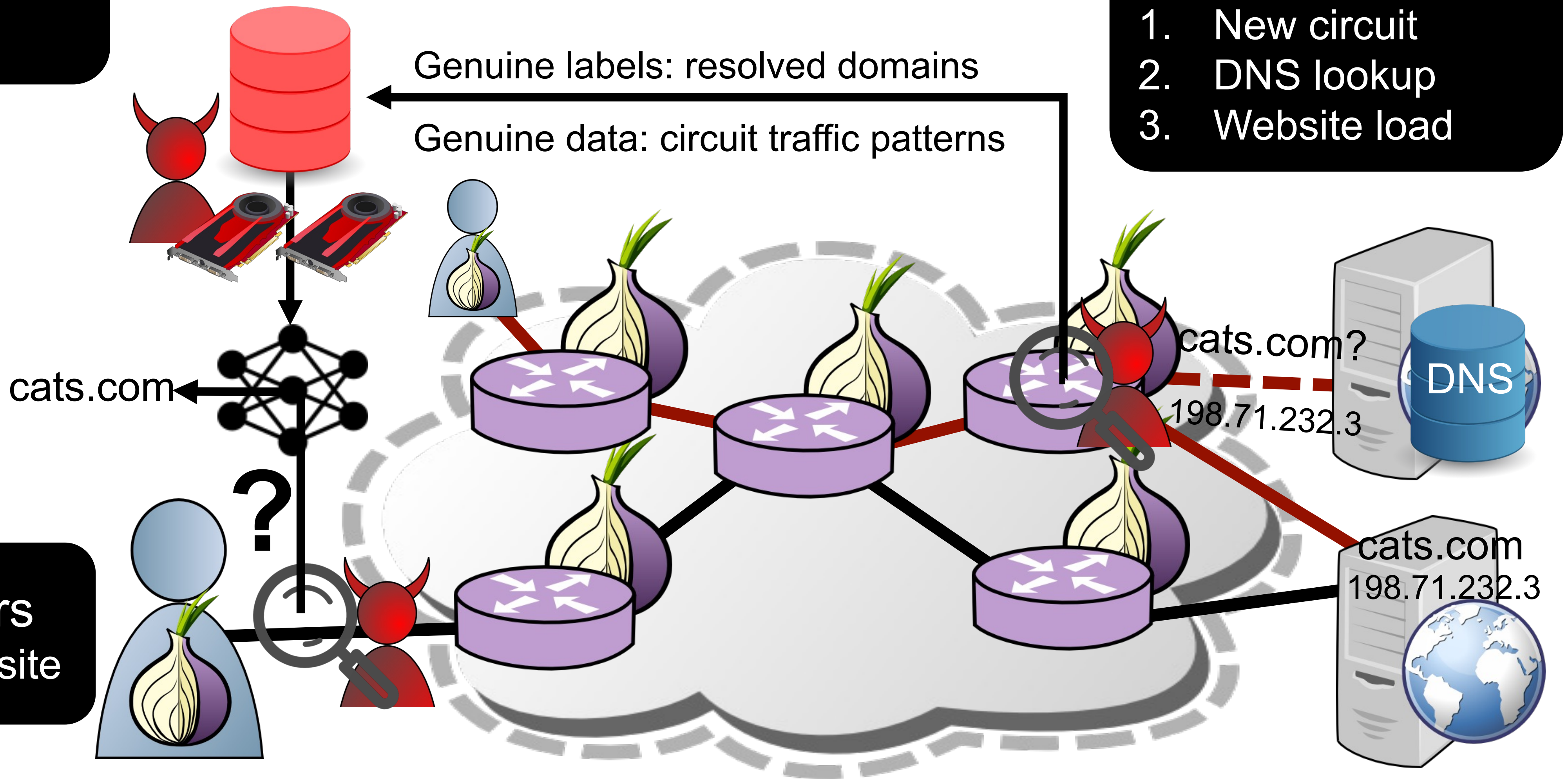
Key Insight: Exits Observe Genuine Data & Labels

Step 1: gather data & labels
- Run a Tor exit relay and use to to collect genuine Tor traffic

Step 2: train ML classifier
- Use collected data & labels

Step 3: deploy against users
- Use ML model to predict website

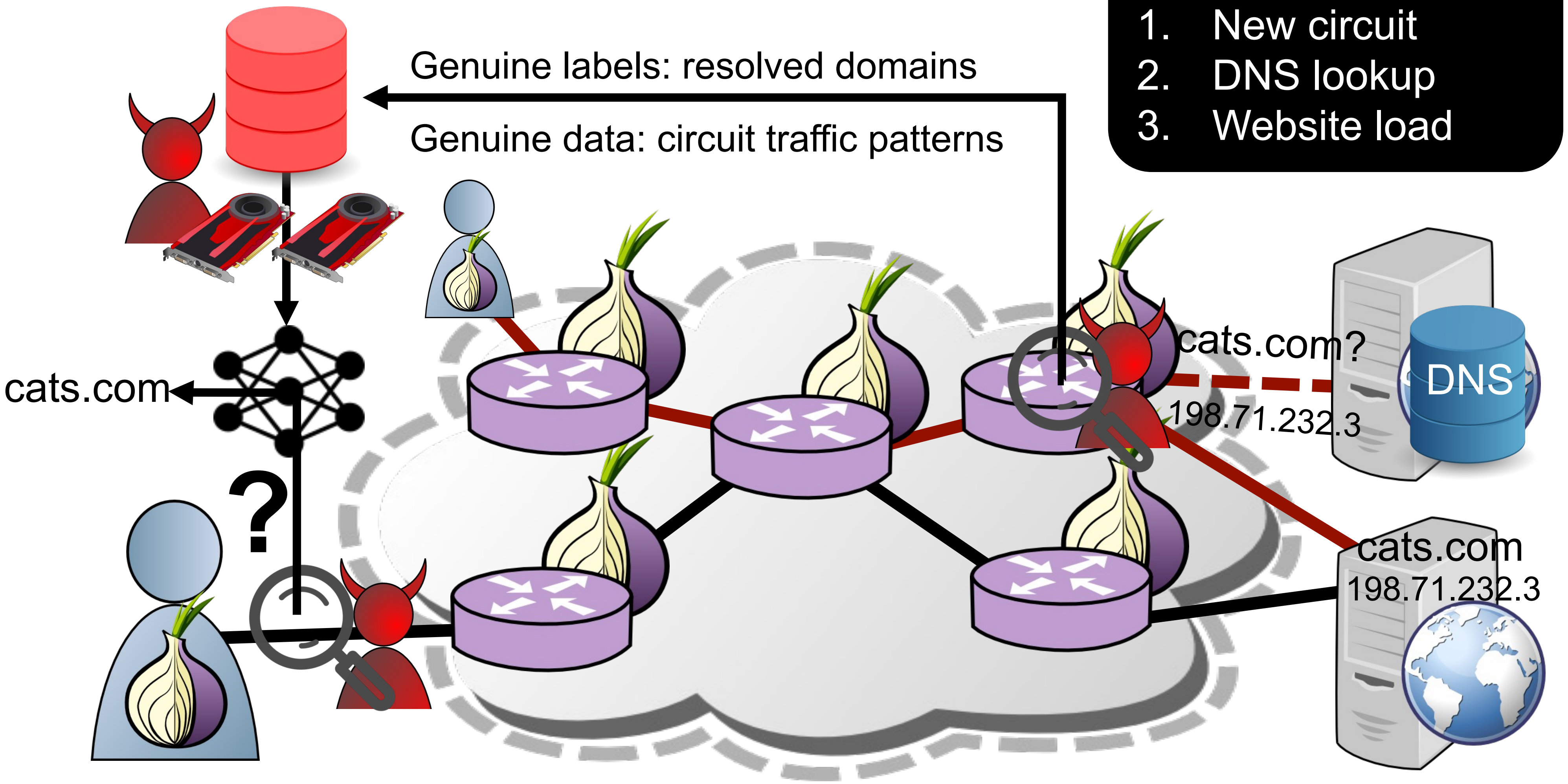
Exit can observe:
1. New circuit
2. DNS lookup
3. Website load



Key Insight: Exits Observe Genuine Data & Labels

Benefits

- Captures real world diversity of browsers, behavior, world size, choice of pages
- Can stop trying to fix the synthetic model



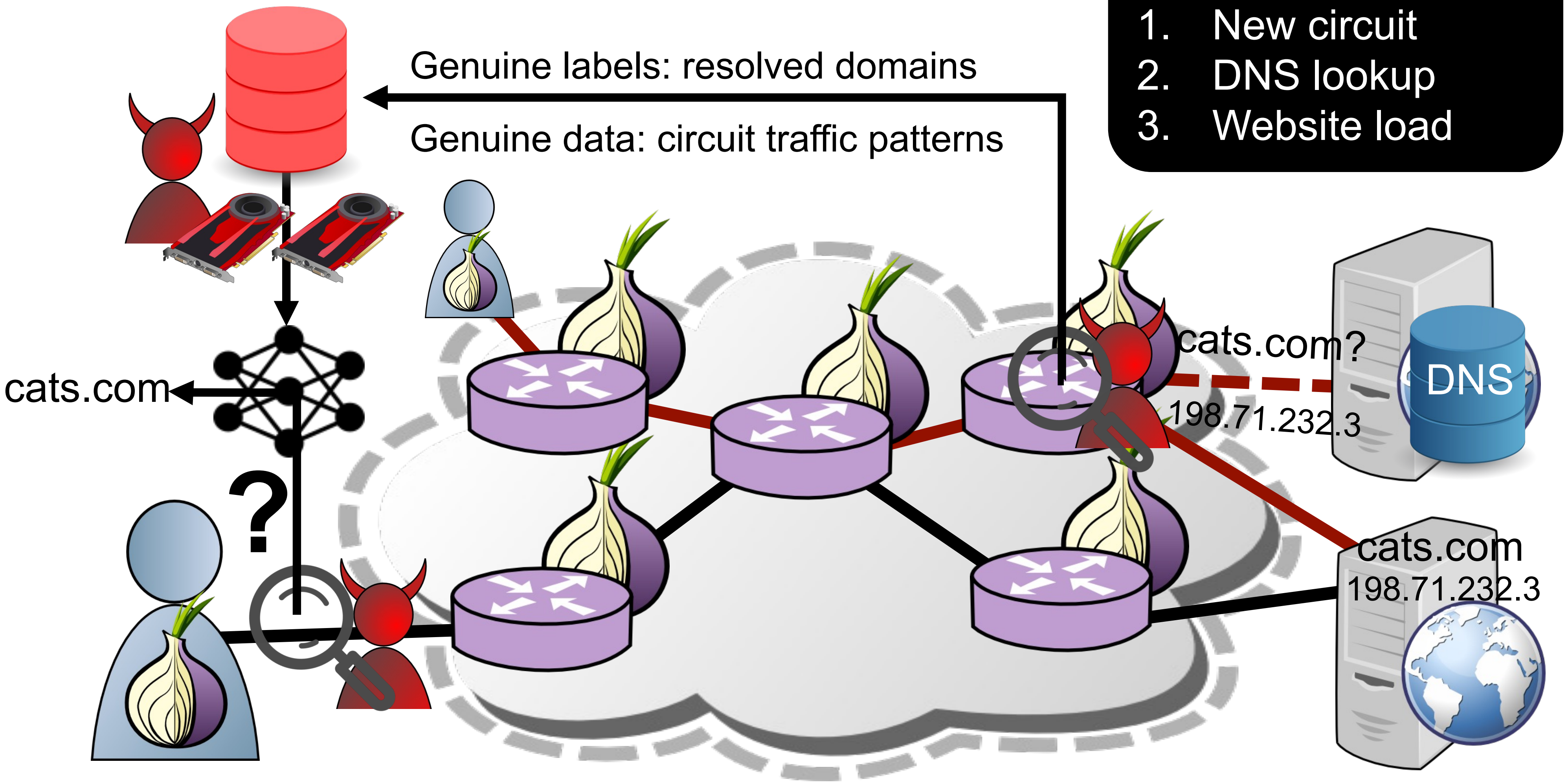
Exit can observe:

1. New circuit
2. DNS lookup
3. Website load

Key Insight: Exits Observe Genuine Data & Labels

Benefits
- Captures real world diversity of browsers, behavior, world size, choice of pages
- Can stop trying to fix the synthetic model

Caveats
- Train at exit, deploy at entry → noise
- Domain, not page label
- Need safe eval methods

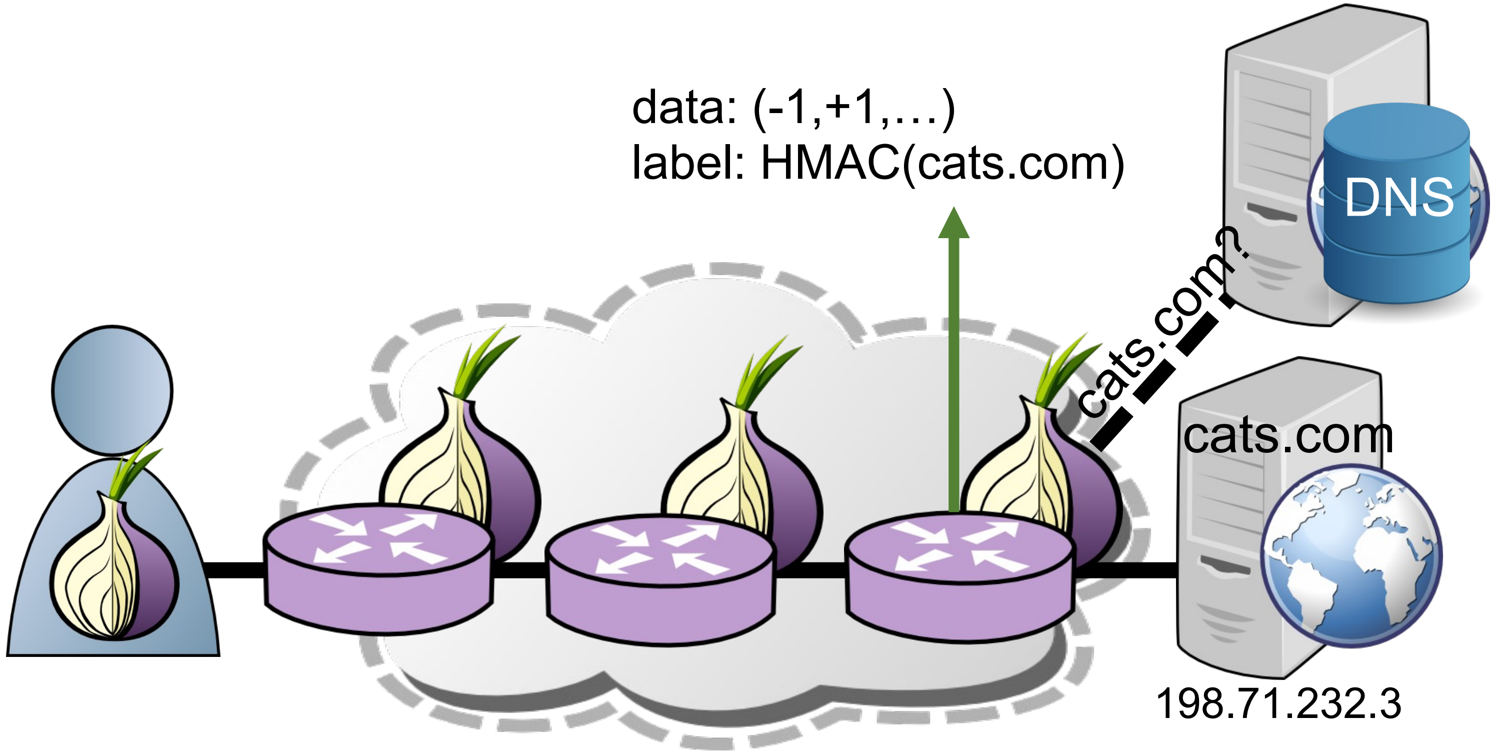


Exit can observe:
1. New circuit
2. DNS lookup
3. Website load

Safe Evaluation using Online Learning

Our safe evaluation plan:

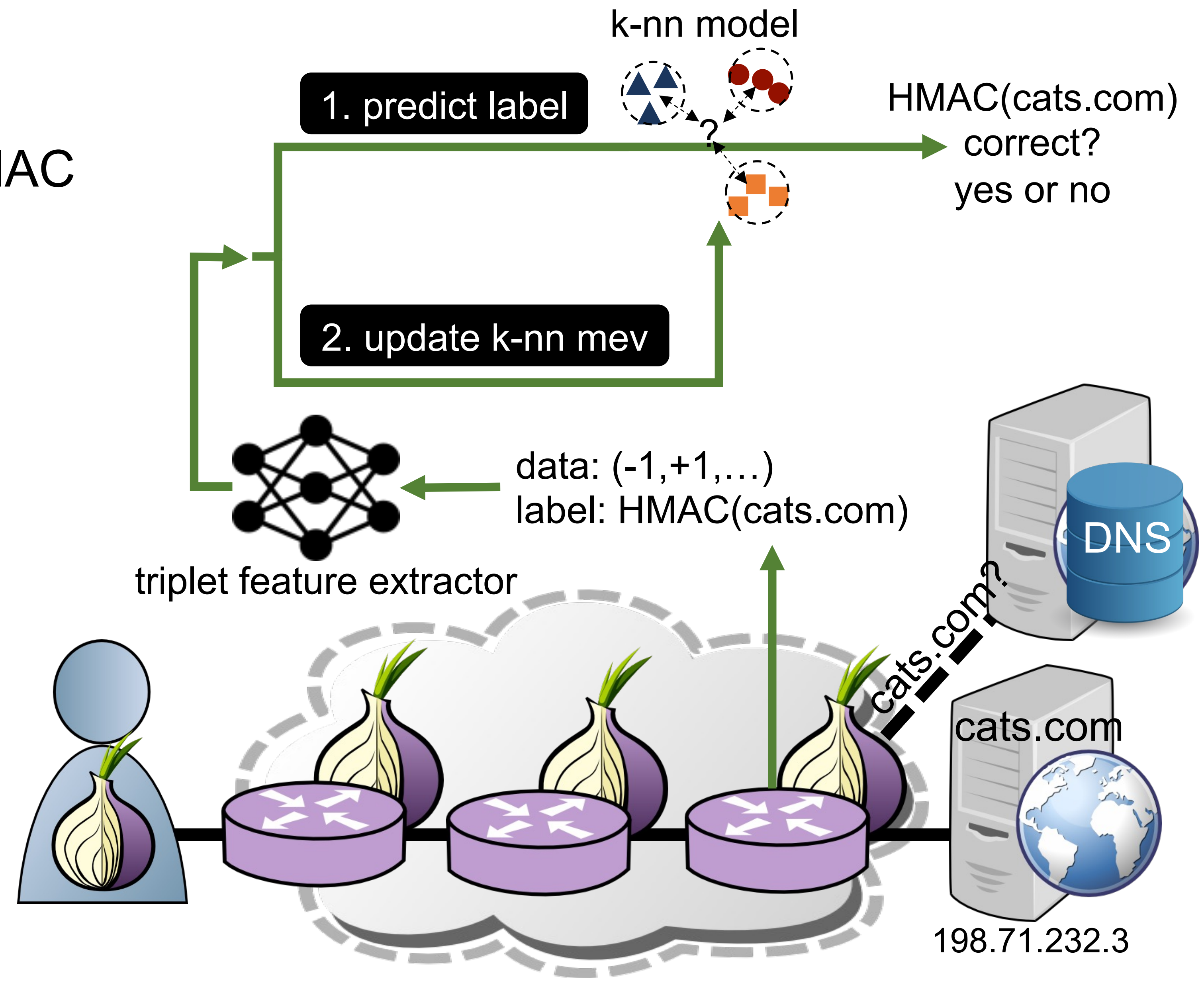
- Hash domain labels using keyed HMAC
  - Never learn true labels



Safe Evaluation using Online Learning

Our safe evaluation plan:

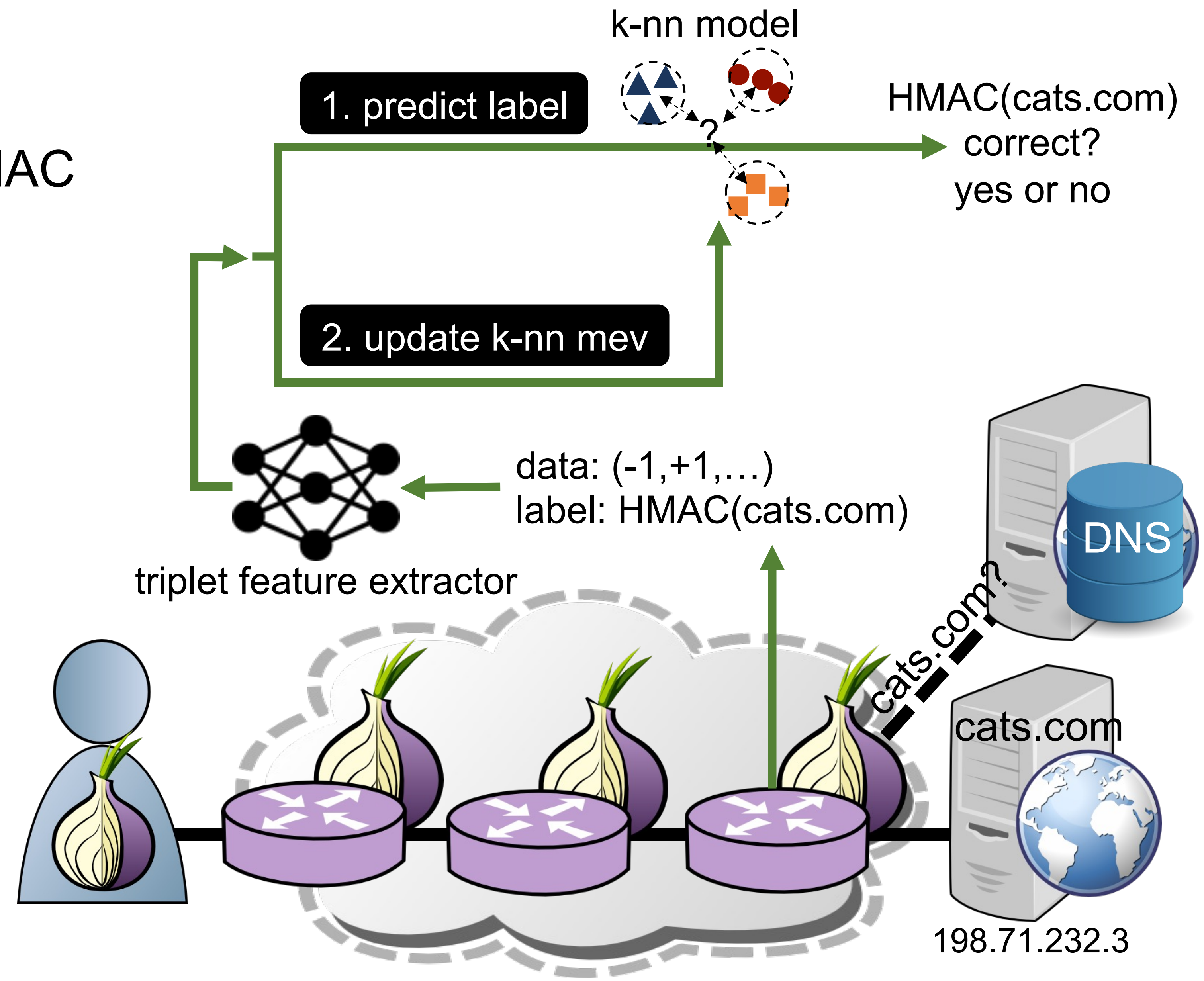
- Hash domain labels using keyed HMAC
  - Never learn true labels
- Use online learning
  - Adapted Triplet Fingerprinting [CCS'19]
  - Compute means in real time, discard data
  - Individual data items never stored



Safe Evaluation using Online Learning

Our safe evaluation plan:

- Hash domain labels using keyed HMAC
  - Never learn true labels
- Use online learning
  - Adapted Triplet Fingerprinting [CCS'19]
  - Compute means in real time, discard data
  - Individual data items never stored
- Other safety precautions
  - Never deanonymizes Tor users
  - Destroyed models, HMAC key after eval
- Tor Safety Board reviewed plan
  - See paper for details!



U.S. NAVAL RESEARCH LABORATORY

Evaluation

Train and evaluate at exit relay

- No noise from transferring to entry
- Upper bound on attack accuracy

Details

- 1 week evaluation
  - 3.9M data sequences, 671k unique sites
- Multi-class classification
  - predict a monitored site, or ‘unmonitored’
- Performance metric
  - instant accuracy (i.e., moving average)
  - # correct / # total predictions (10k window)

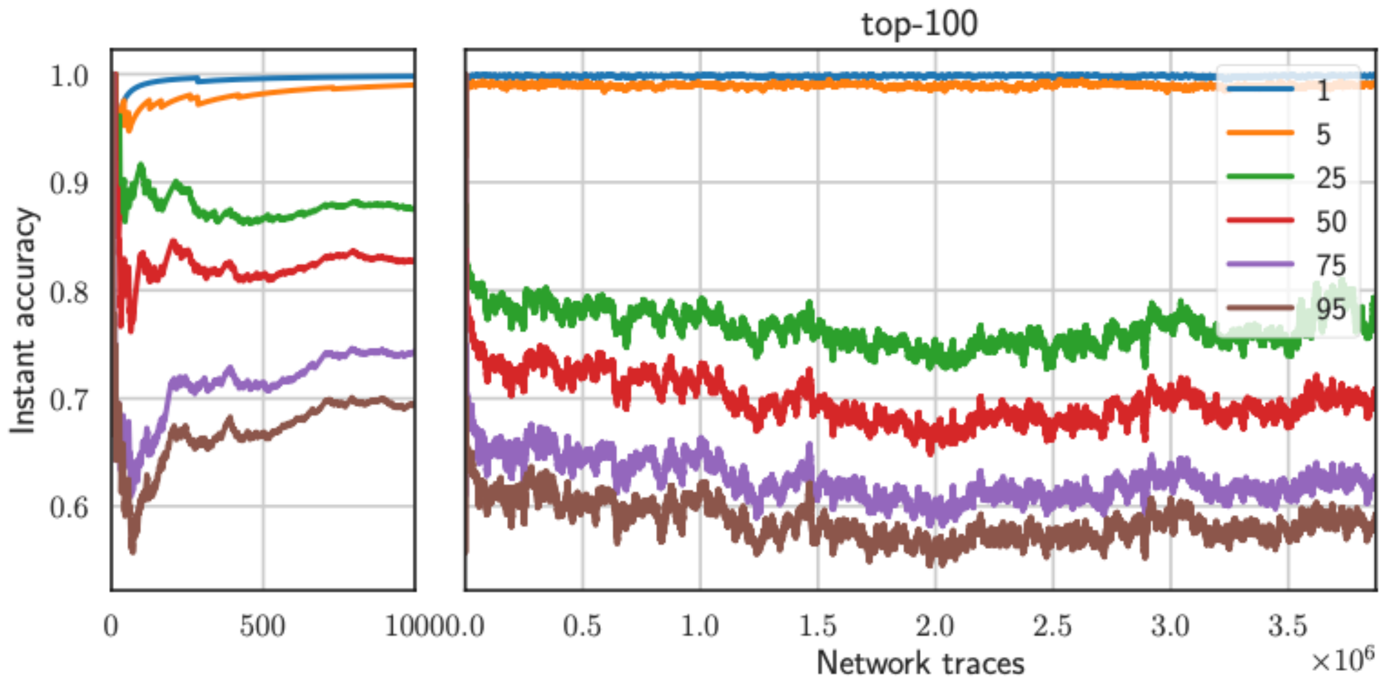
Evaluation

Train and evaluate at exit relay

- No noise from transferring to entry
- Upper bound on attack accuracy

Details

- 1 week evaluation
  - 3.9M data sequences, 671k unique sites
- Multi-class classification
  - predict a monitored site, or ‘unmonitored’
- Performance metric
  - instant accuracy (i.e., moving average)
  - # correct / # total predictions (10k window)



Evaluation

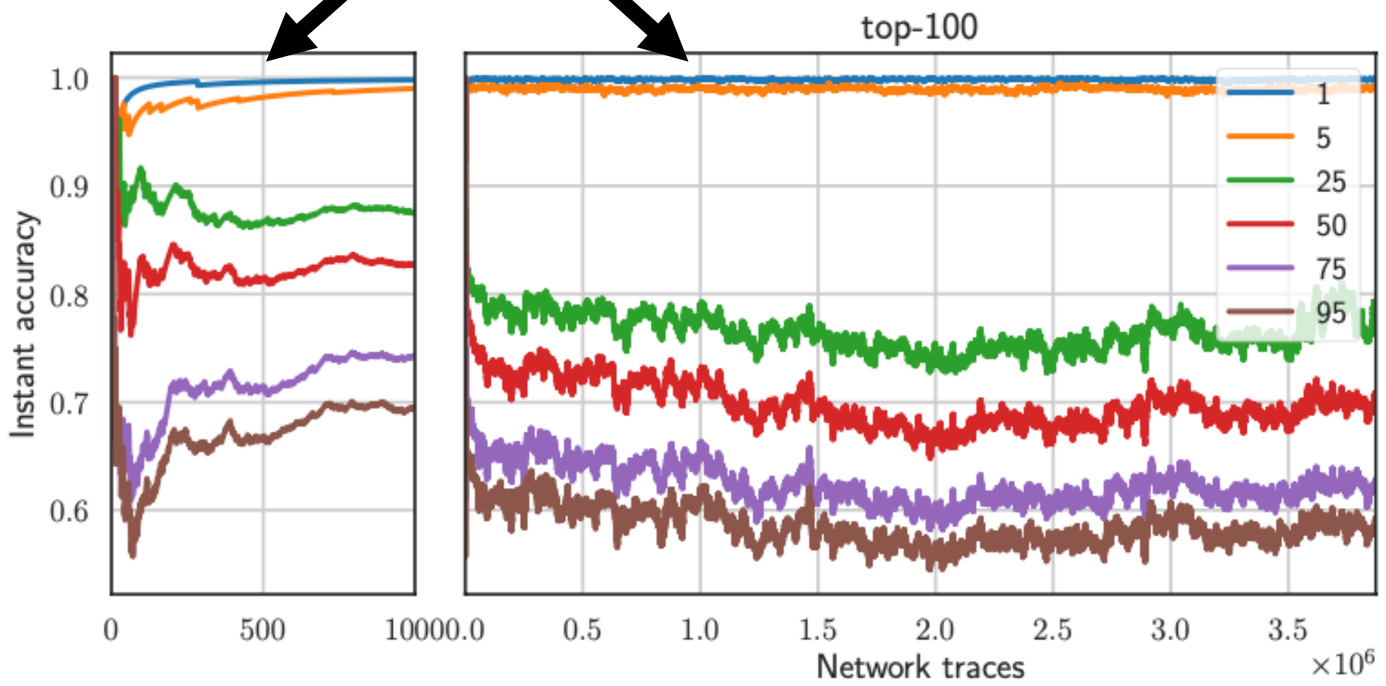
Train and evaluate at exit relay

- No noise from transferring to entry
- Upper bound on attack accuracy

Details

- 1 week evaluation
  - 3.9M data sequences, 671k unique sites
- Multi-class classification
  - predict a monitored site, or ‘unmonitored’
- Performance metric
  - instant accuracy (i.e., moving average)
  - # correct / # total predictions (10k window)

accuracy above 95% when monitoring ≤ 5 sites



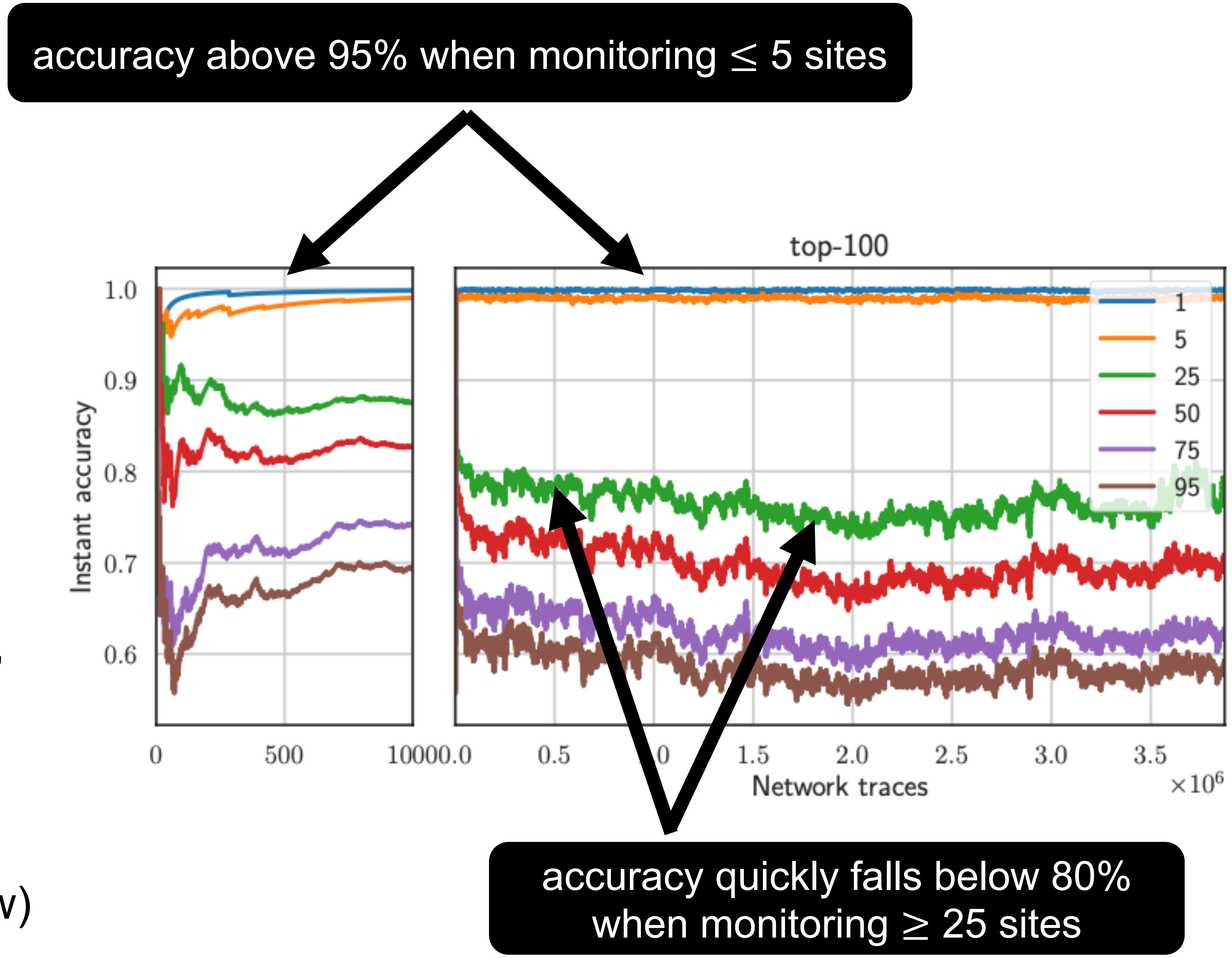
Evaluation

Train and evaluate at exit relay

- No noise from transferring to entry
- Upper bound on attack accuracy

Details

- 1 week evaluation
  - 3.9M data sequences, 671k unique sites
- Multi-class classification
  - predict a monitored site, or 'unmonitored'
- Performance metric
  - instant accuracy (i.e., moving average)
  - # correct / # total predictions (10k window)



Genuine vs. Synthetic Data

Offline phase

- Crawl ‘synthetic’ list of domains
  - Synthetic: use crawl to train a classifier offline

Online phase

- Train two classifiers online
  - Hybrid: update copy of synthetic classifier with genuine data
  - Real: train new classifier on genuine data only

- 1 week evaluation
  - 1.2M data sequences
  - observed 183 of 1,074 ‘synthetic’ domains

- Binary classification
  - monitored set contains 5 sites
  - predict either ‘monitored’ or ‘unmonitored’

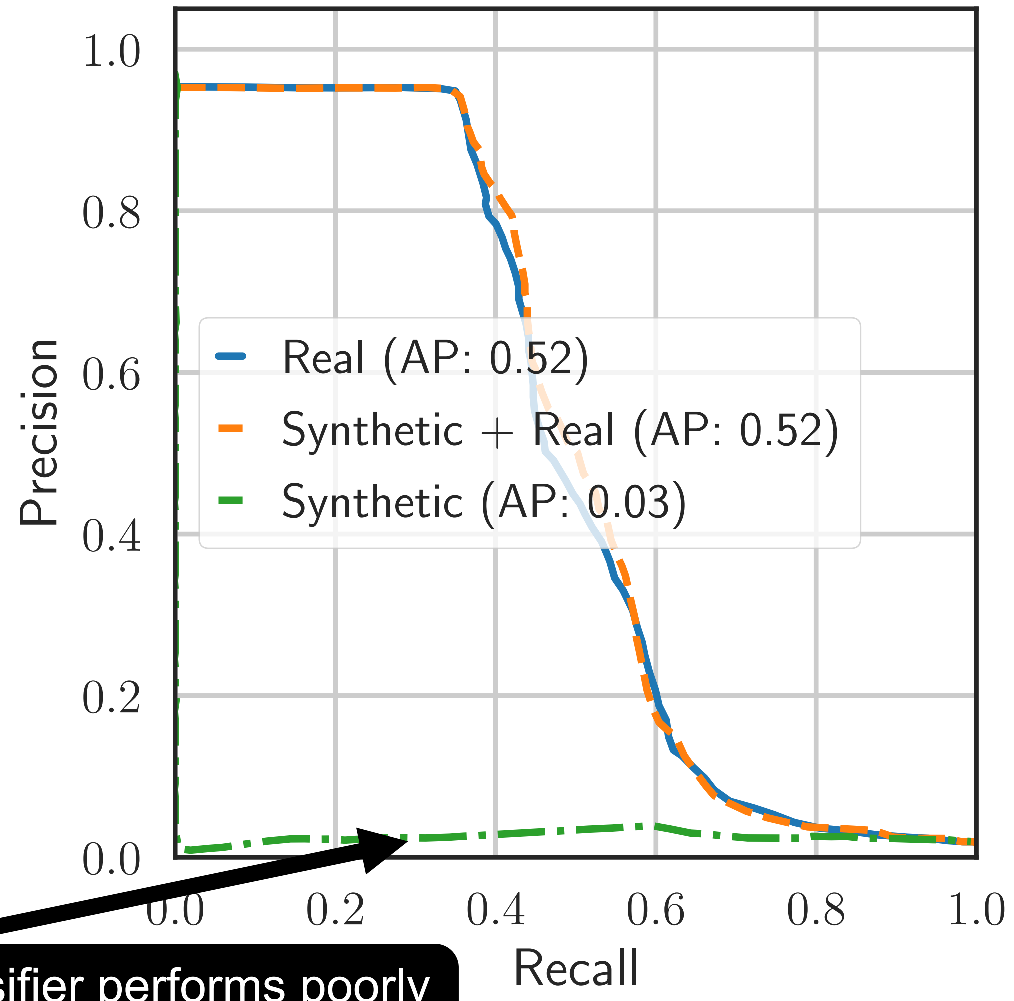
Genuine vs. Synthetic Data

Offline phase

- Crawl ‘synthetic’ list of domains
  - Synthetic: use crawl to train a classifier offline

Online phase

- Train two classifiers online
  - Hybrid: update copy of synthetic classifier with genuine data
  - Real: train new classifier on genuine data only
- 1 week evaluation
  - 1.2M data sequences
  - observed 183 of 1,074 ‘synthetic’ domains
- Binary classification
  - monitored set contains 5 sites
  - predict either ‘monitored’ or ‘unmonitored’



synthetic classifier performs poorly against genuine data

Genuine vs. Synthetic Data

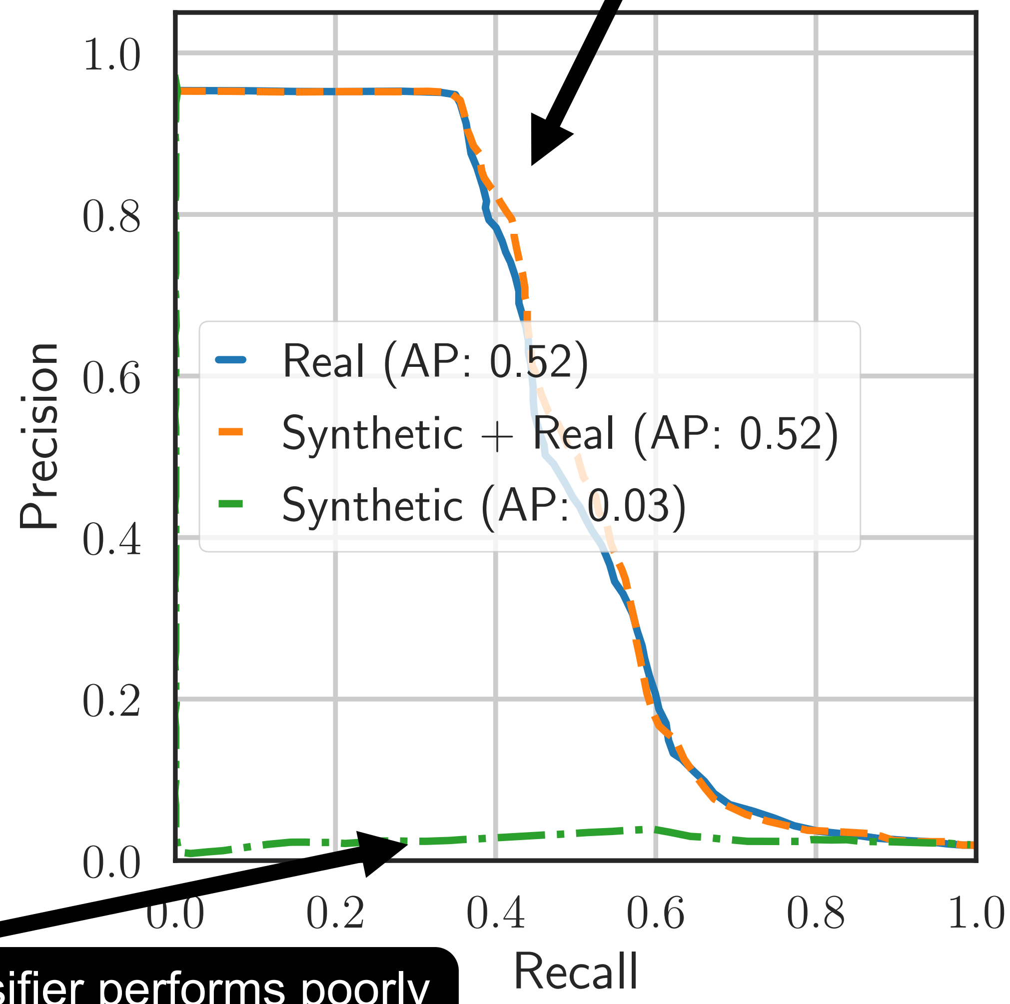
Offline phase

- Crawl ‘synthetic’ list of domains
  - Synthetic: use crawl to train a classifier offline

Online phase

- Train two classifiers online
  - Hybrid: update copy of synthetic classifier with genuine data
  - Real: train new classifier on genuine data only
- 1 week evaluation
  - 1.2M data sequences
  - observed 183 of 1,074 ‘synthetic’ domains
- Binary classification
  - monitored set contains 5 sites
  - predict either ‘monitored’ or ‘unmonitored’

synthetic data does not improve model over genuine data



synthetic classifier performs poorly against genuine data

U.S. NAVAL RESEARCH LABORATORY

Training and Testing on Opposite Ends

Fully synthetic evaluation

- Crawled 1k URLs 10x each
- Pinned entry and exit on each circuit
- Collected data sequences in both positions on each circuit
- Closed-world batch classification
  - 50%-50% train-test split

Monitored set size:	**5**	**50**	**750**
Train and test on <u>exit</u>	91.2%	76.2%	52.2%
Train on <u>exit</u>, test on <u>entry</u>	86.4%	65.1%	34.1%
Loss in accuracy:	**4.8%**	**11.1%**	**18.1%**

loss in accuracy is low for feasible (i.e. small) monitored sets

Main Takeaways

Insights

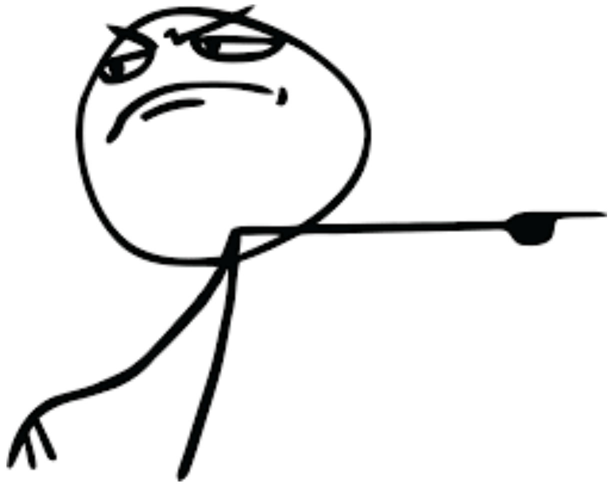
- WF can be feasible with genuine data and small monitored sets, online learning can mitigate concept drift
- Synthetic data is not useful when the adversary deploys in the real world
- Simple defenses may be more effective than we thought
  - Adversary has to simulate defense on top of undefended exit data

Future Research Areas

- Improve accuracy when training on genuine data
- Reduce distortion when transferring models from exit to entry
- Defenses that make it harder to learn from genuine data, increase distortion

Contact

- rob.g.jansen@nrl.navy.mil
- robgjansen.com
- @robgjansen



Read the paper!