U.S. NAVAL RESEARCH LABORATORY

Repositioning Real-World Website Fingerprinting on Tor with a Measurement of Genuine Tor Traces

Rob Jansen, U.S. Naval Research Laboratory
Ryan Wails, U.S. Naval Research Laboratory and Georgetown University
Aaron Johnson, U.S. Naval Research Laboratory

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

Cyber Security Seminar
University of Edinburgh
March 10th, 2025

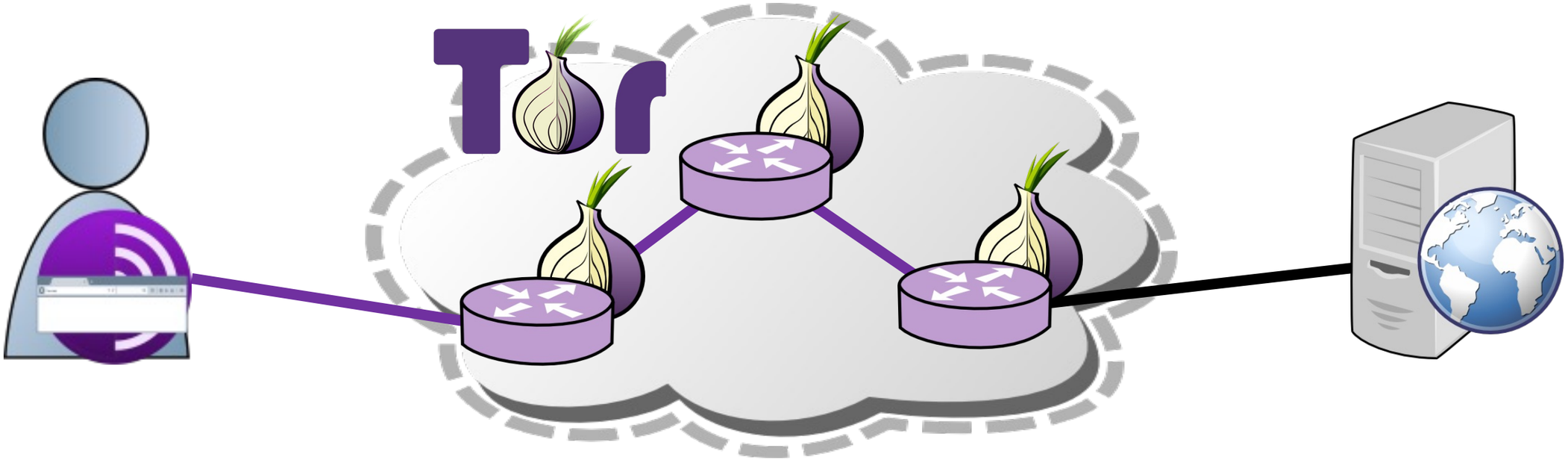
U.S. NAVAL RESEARCH LABORATORY

Anonymous Communication with Tor

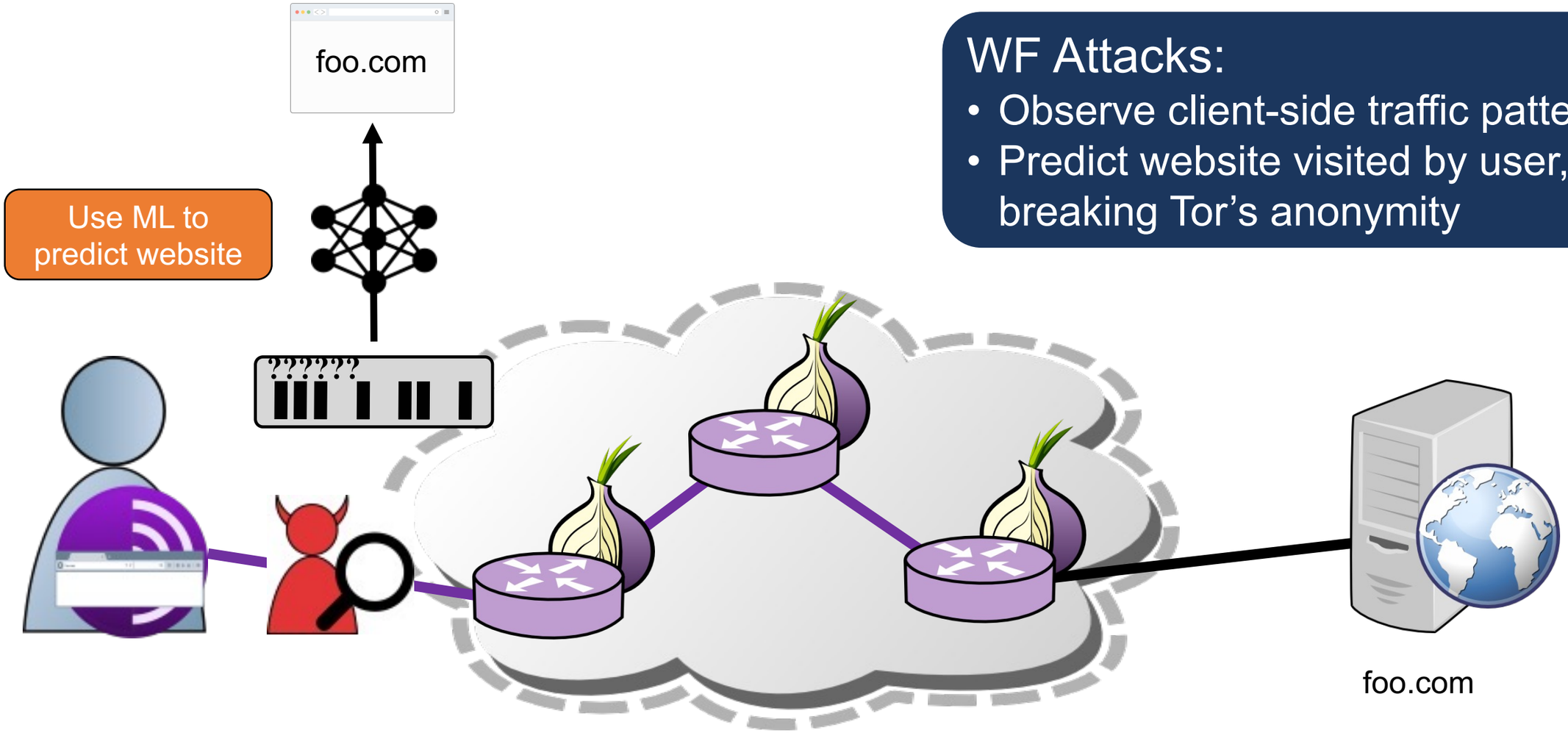
- Separates *identification* from *routing*
- Provides unlinkable communication
- Promotes user safety and privacy online

Tor Browse Privately.
Explore Freely.

Defend yourself against tracking and surveillance. Circumvent censorship.



Website Fingerprinting (WF) Threat Model

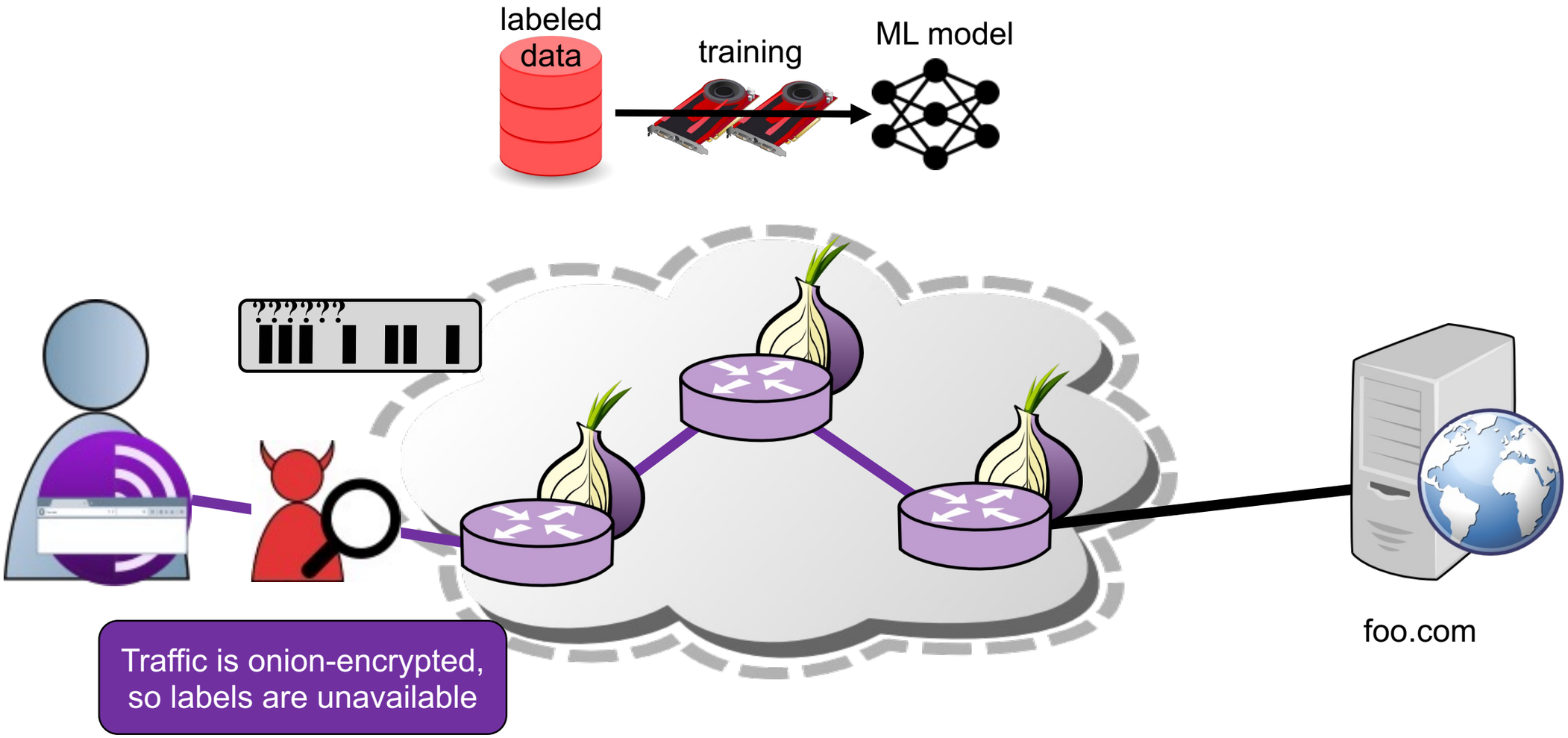


WF Attacks:

- Observe client-side traffic patterns
- Predict website visited by user, breaking Tor's anonymity

U.S. NAVAL RESEARCH LABORATORY

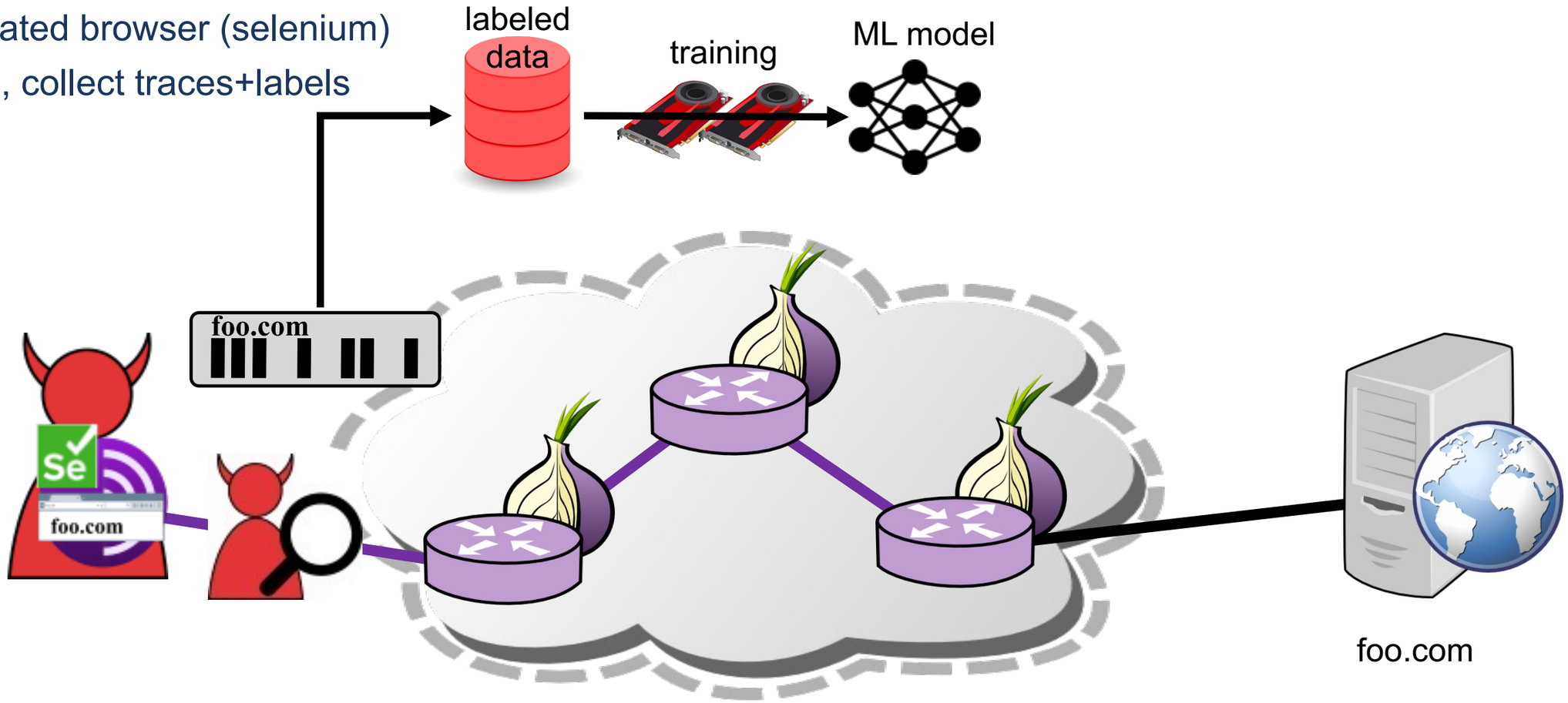
How Might an Adversary Train its ML Models?



How Might an Adversary Train its ML Models?

Traditional method?

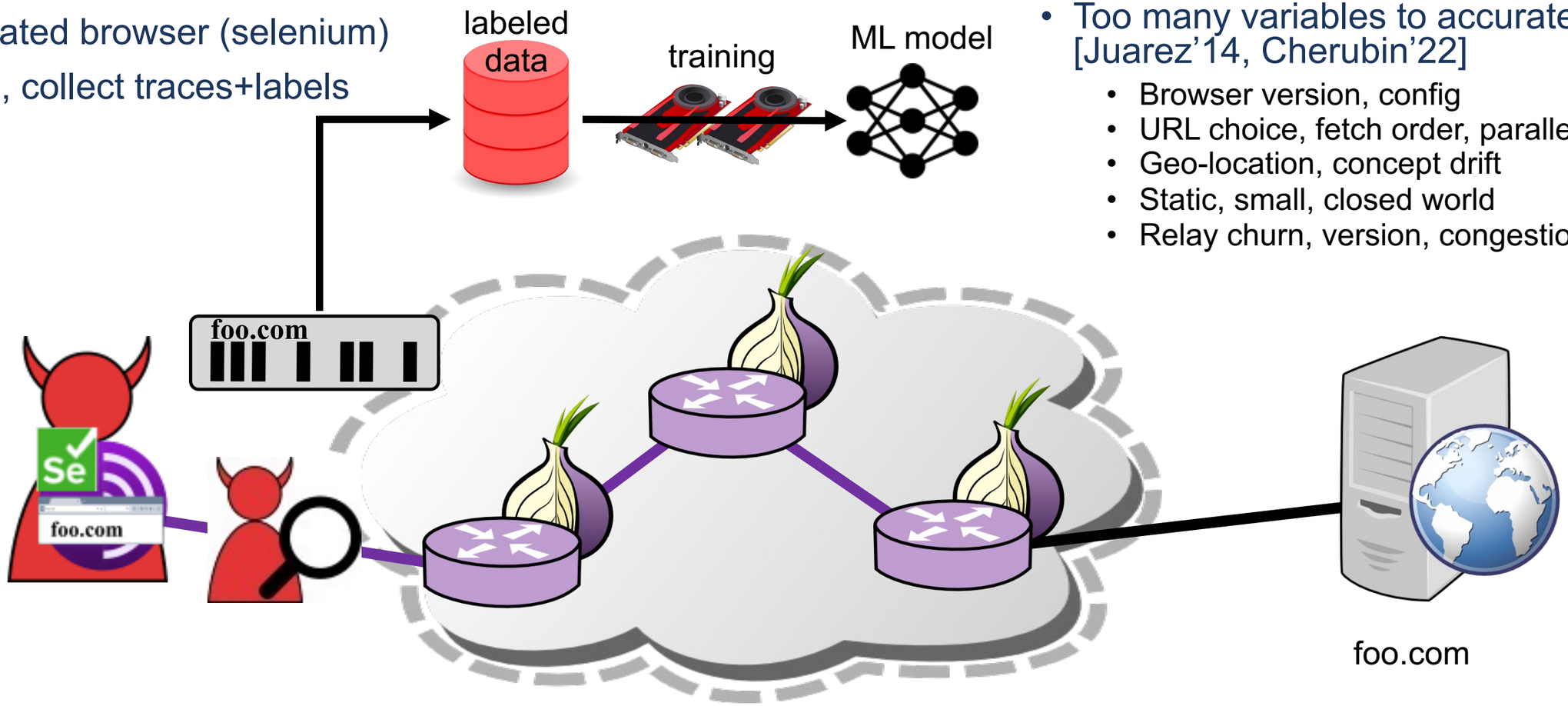
- Use automated browser (selenium)
- Crawl sites, collect traces+labels



How Might an Adversary Train its ML Models?

Traditional method?

- Use automated browser (selenium)
- Crawl sites, collect traces+labels



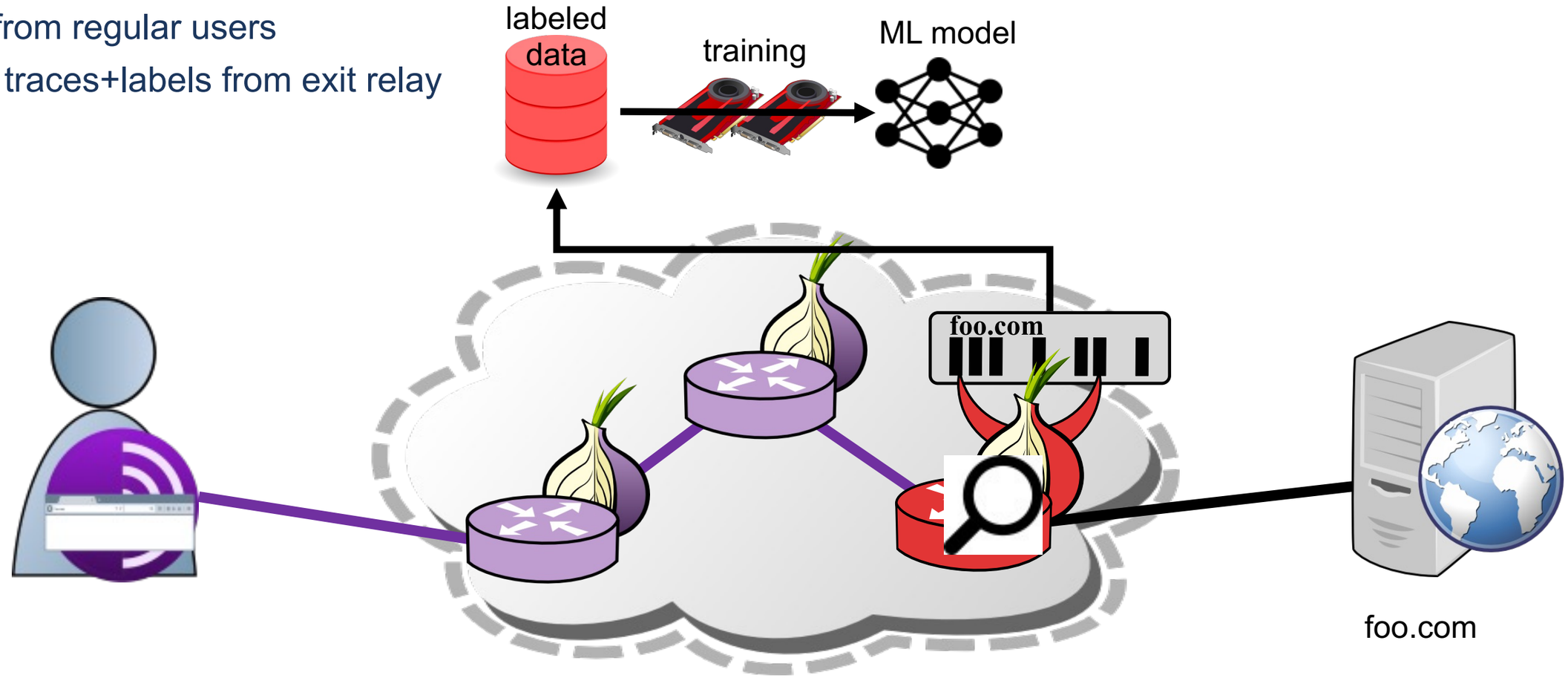
Problems:

- Too many variables to accurately model [Juarez'14, Cherubin'22]
  - Browser version, config
  - URL choice, fetch order, parallel tabs
  - Geo-location, concept drift
  - Static, small, closed world
  - Relay churn, version, congestion, etc.

How Might an Adversary Train its ML Models?

Emerging exit method?

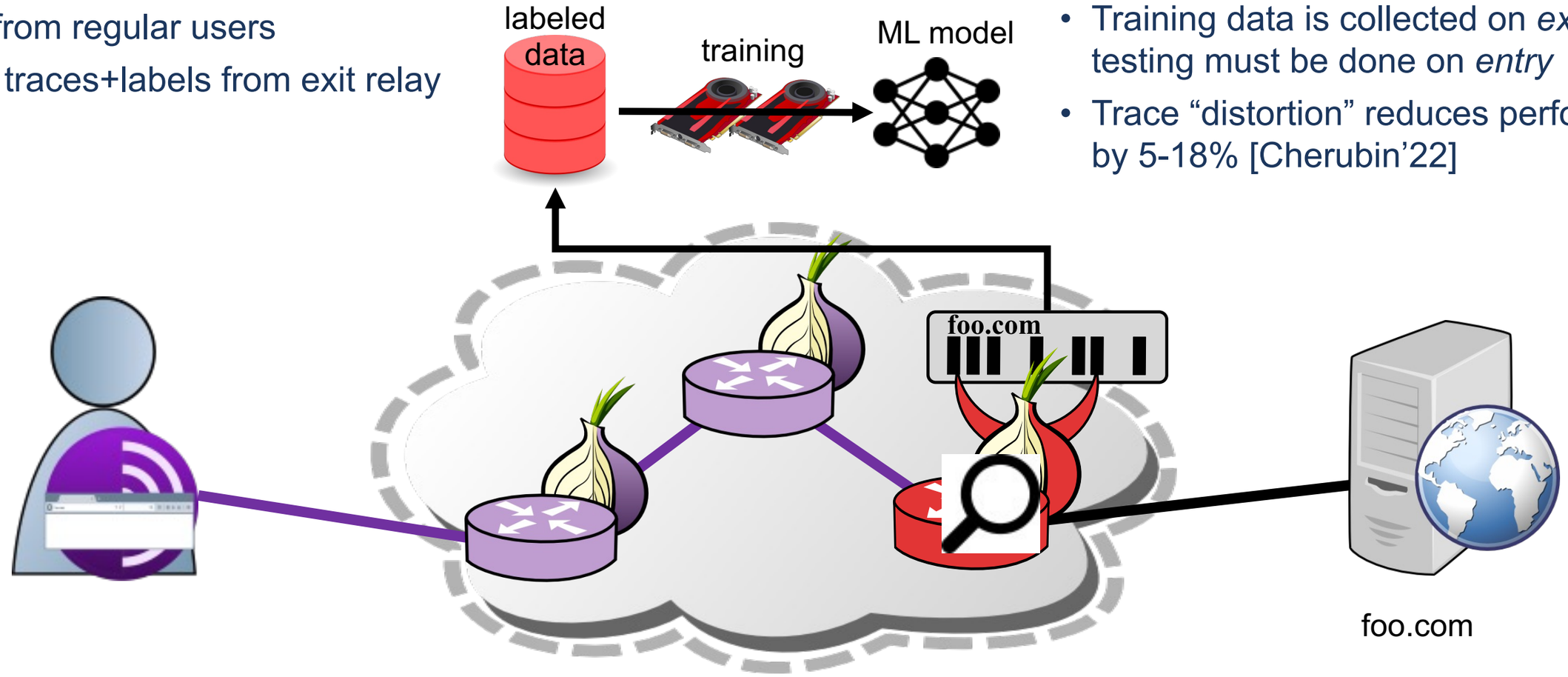
- Traffic from regular users
- Collect traces+labels from exit relay



How Might an Adversary Train its ML Models?

Emerging exit method?

- Traffic from regular users
- Collect traces+labels from exit relay



Problems:

- Training data is collected on *exit*, but testing must be done on *entry*
- Trace "distortion" reduces performance by 5-18% [Cherubin'22]

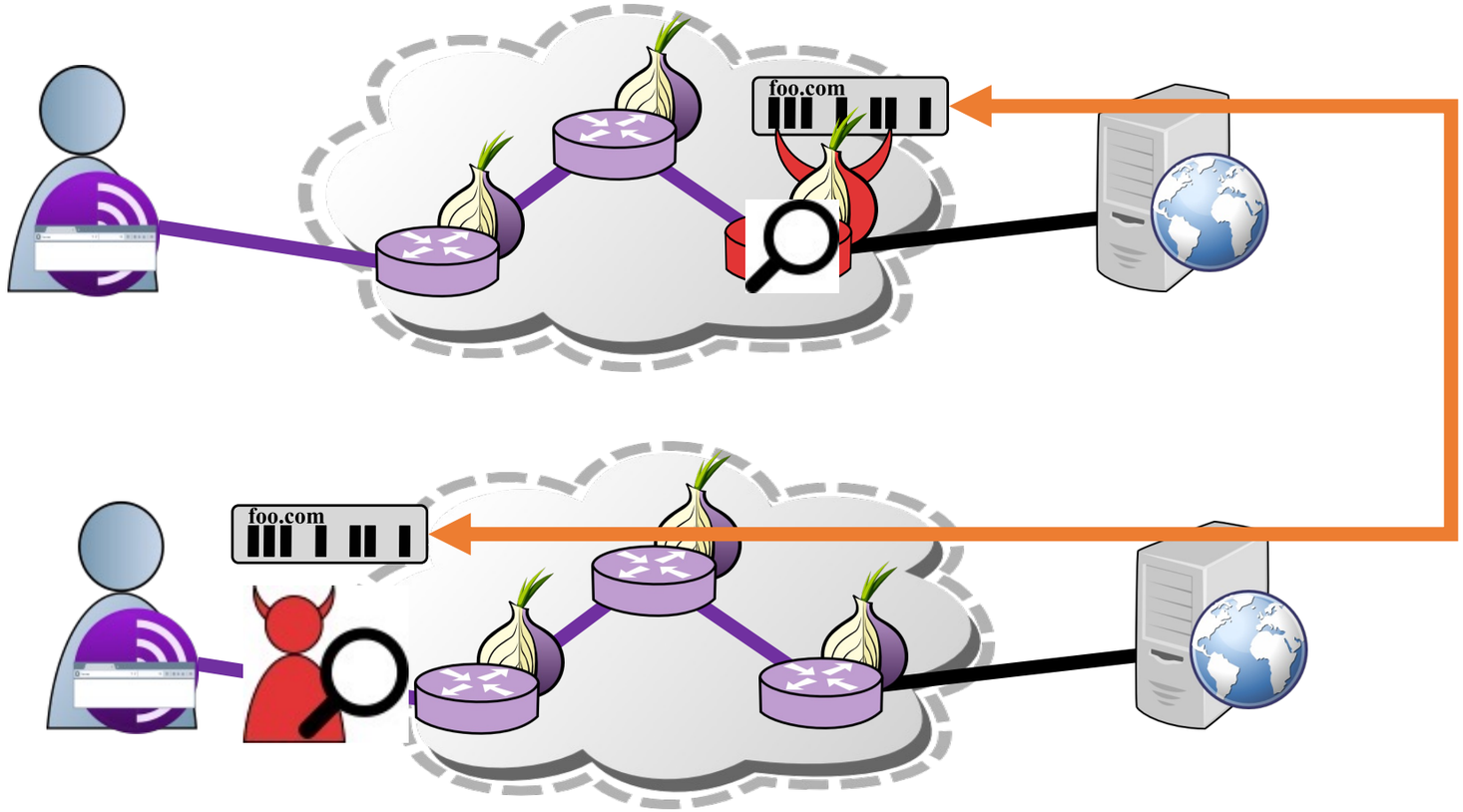
Research Direction

Research Question:

- How can we mitigate trace distortion so that we can utilize real-world traces to better estimate the threat of WF against Tor?

Training

Testing



Mitigate distortion between traces from entry and exit positions

Outline

1. Trace transduction with Retracer
2. Retracer evaluation
3. Real-world WF evaluation

Cell Trace Transduction

Cell Trace Transduction

- Cell trace:
  - a sequence of *n* (*timestamp*, *direction*) pairs
    - timestamp: when cell was observed, relative to start of connection
    - direction: +1 if forwarded toward server, -1 if toward client

Example cell trace:

```
[
   (0.1, +1),
   (0.5, -1),
   (0.9, +1),
   (1.3, -1),
   (1.3, -1),
   (1.3, -1),
   …
]
```

Cell Trace Transduction

Cell Trace Transduction

- Cell trace:
  - a sequence of *n* (*timestamp*, *direction*) pairs
    - timestamp: when cell was observed, relative to start of connection
    - direction: +1 if forwarded toward server, -1 if toward client
- Transducer:
  - a function $T(I, M, p_{in}, p_{out}) \rightarrow [O]_M$
  - transforms an input cell trace *I* in position p_{in} into *M* output cell traces *O* in position p_{out}
  - we want p_{in}=exit, p_{out}=entry

Example cell trace:

```
[
   (0.1, +1),
   (0.5, -1),
   (0.9, +1),
   (1.3, -1),
   (1.3, -1),
   (1.3, -1),
   …
]
```

U.S. NAVAL RESEARCH LABORATORY

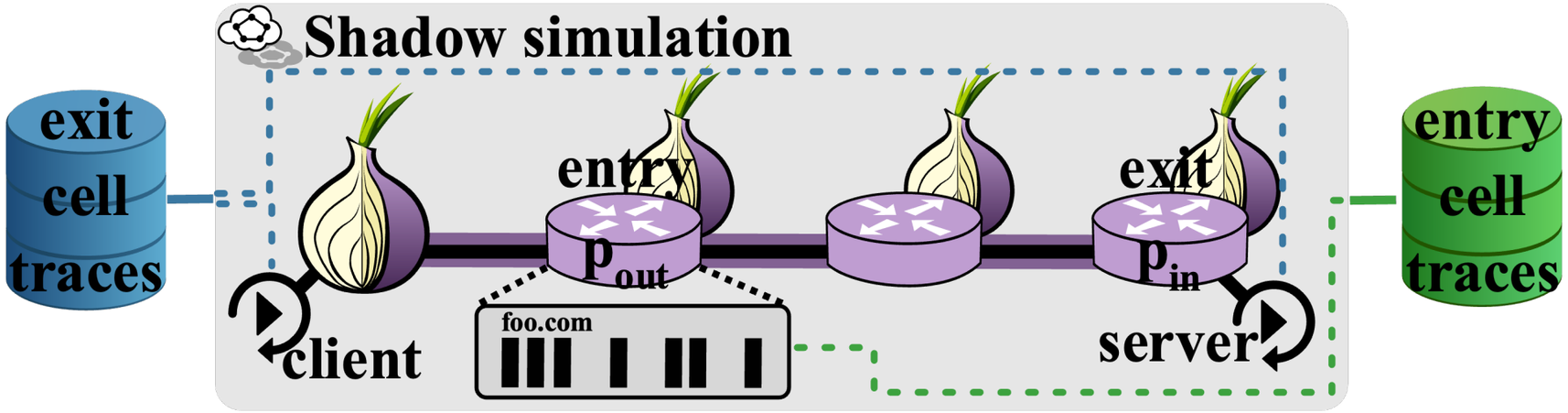
Retracer: A Cell Trace Transducer

- Key Insights
  - A cell trace has the metadata needed to reproduce it
  - Network simulation tools (Shadow) model Tor with high fidelity
  - We can replay an *exit* trace in Shadow and extract its *entry* trace

U.S. NAVAL RESEARCH LABORATORY

Retracer: A Cell Trace Transducer

- Key Insights
  - A cell trace has the metadata needed to reproduce it
  - Network simulation tools (Shadow) model Tor with high fidelity
  - We can replay an *exit* trace in Shadow and extract its *entry* trace
- Retracer
  - Replays cells traces in large-scale Tor simulations with Shadow
  - Uses cell trace timing and directions as a transcript for replay
  - Adjusts for latency between client and exit during replay

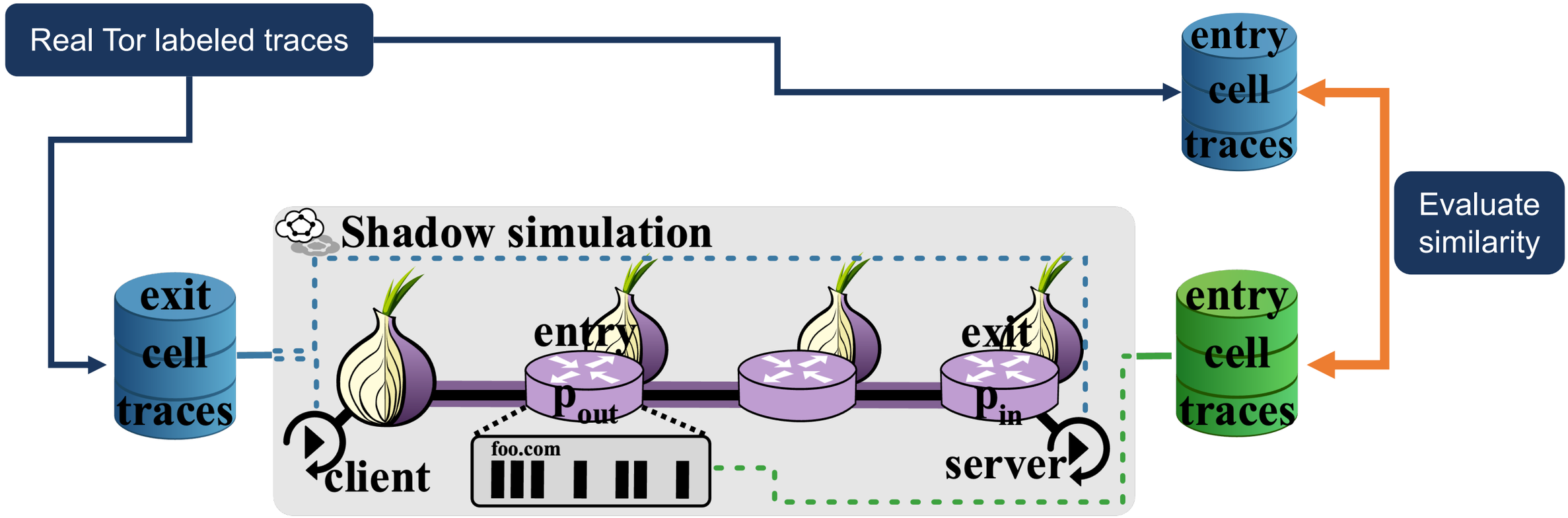


Outline

1. Trace transduction with Retracer
2. Retracer evaluation
3. Real-world WF evaluation

Retracer Evaluation Plan

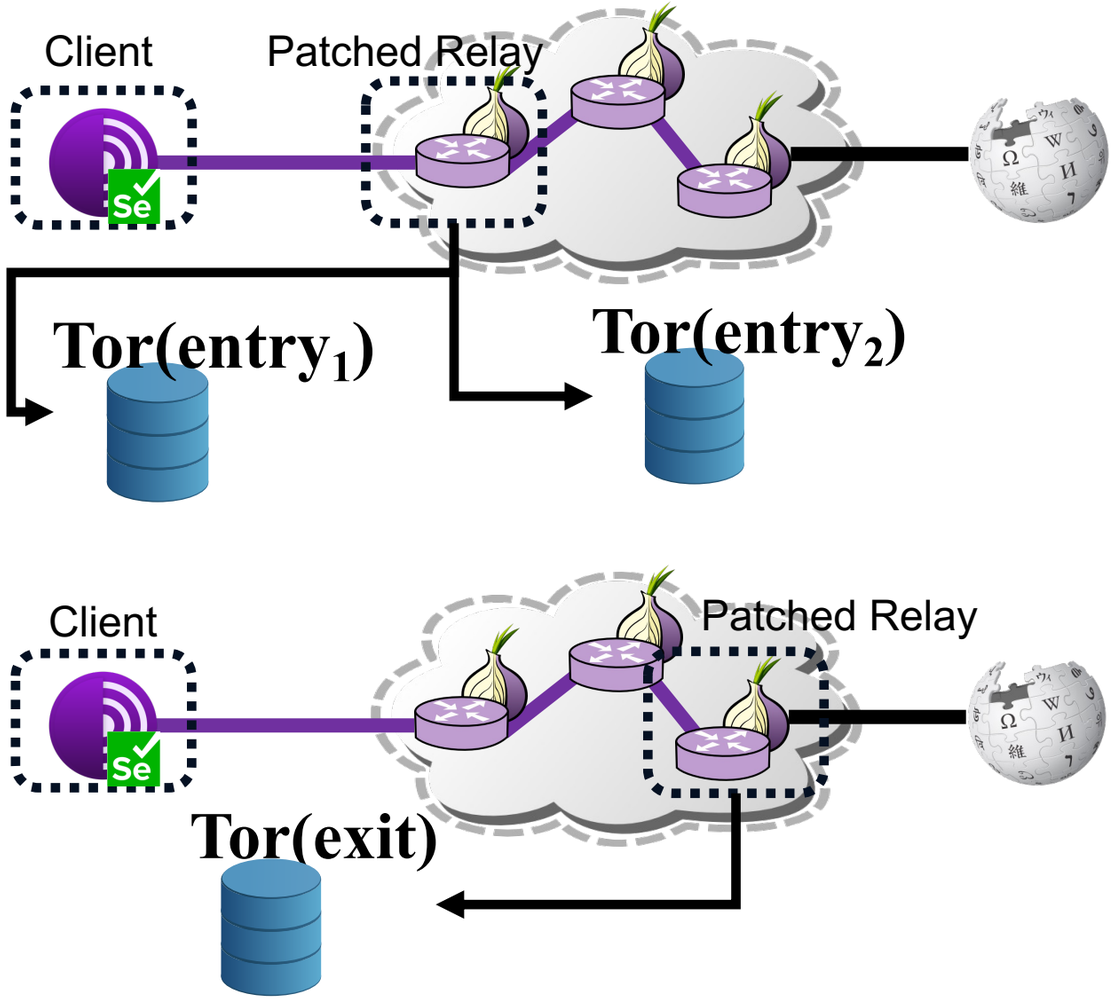
Goal: evaluate how well Retracer transduces *exit* to *entry* traces



Collecting datasets for Retracer evaluation

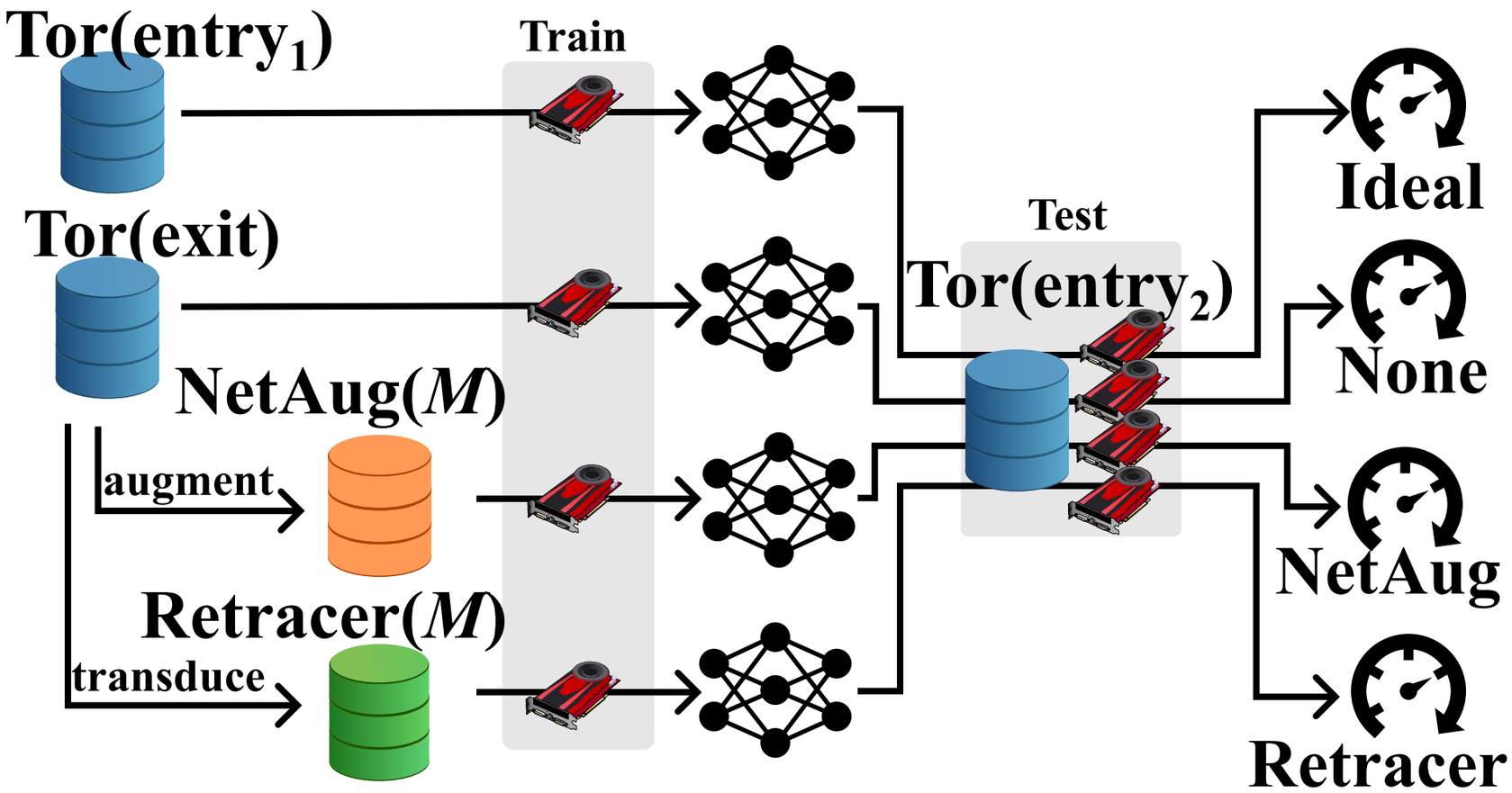
Tor Dataset Collection

- Patch Tor relay to record cell traces (*only* those from *our* client)
- Select some Wikipedia pages
- Fetch each page multiple times through our Tor relay, record traces
- Repeat through Tor exit and entry positions



Retracer Evaluation Methodology

We measure Retracer's efficacy using a downstream WF classification task



U.S. NAVAL RESEARCH LABORATORY

Retracer Evaluation Results

Table 2: Classifier Accuracy in a Multiclass Closed World Classification Experiment when Tested on $\mathrm{Tor}(\mathrm{entry}_2)$

Method	**Training set**	**DF**	**Tik-Tok**
Ideal	$\mathrm{Tor}(\mathrm{entry}_1)$	89%	87%
Retracer	Retracer(19)	86% (↓ 3 pp)	85% (↓2 pp)
NetAug	NetAug(19)	70% (↓19 pp)	⊥
None	Tor(exit)	76% (↓13 pp)	79% (↓8 pp)
Classifier Properties →		Time-Oblivious	Time-Aware

⊥: Timing information required by classifier but unavailable in data.

U.S. NAVAL RESEARCH LABORATORY

Retracer Evaluation Results

Table 2: Classifier Accuracy in a Multiclass Closed World Classification Experiment when Tested on $\mathrm{Tor}(\mathrm{entry}_2)$

Method	Training set	DF	Tik-Tok
Ideal	$\mathrm{Tor}(\mathrm{entry}_1)$	89%	87%
Retracer	Retracer(19)	86% (↓ 3 pp)	85% (↓2 pp)
NetAug	NetAug(19)	70% (↓19 pp)	⊥
None	Tor(exit)	76% (↓13 pp)	79% (↓8 pp)
Classifier Properties →		Time-Oblivious	Time-Aware

⊥: Timing information required by classifier but unavailable in data.

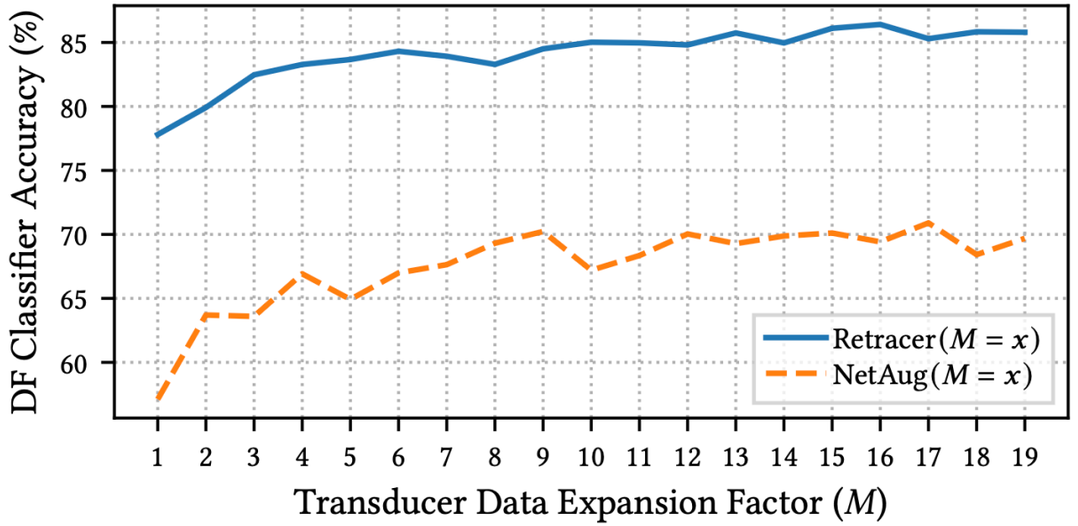


Figure 4: DF classifier accuracy in a multiclass closed-world experiment when training on datasets transduced with an increasing data expansion factor M **and tested on** $\mathrm{Tor}(\mathrm{entry}_2)$.

Outline

1. Trace transduction with Retracer
2. Retracer evaluation
3. Real-world WF evaluation

U.S. NAVAL RESEARCH LABORATORY

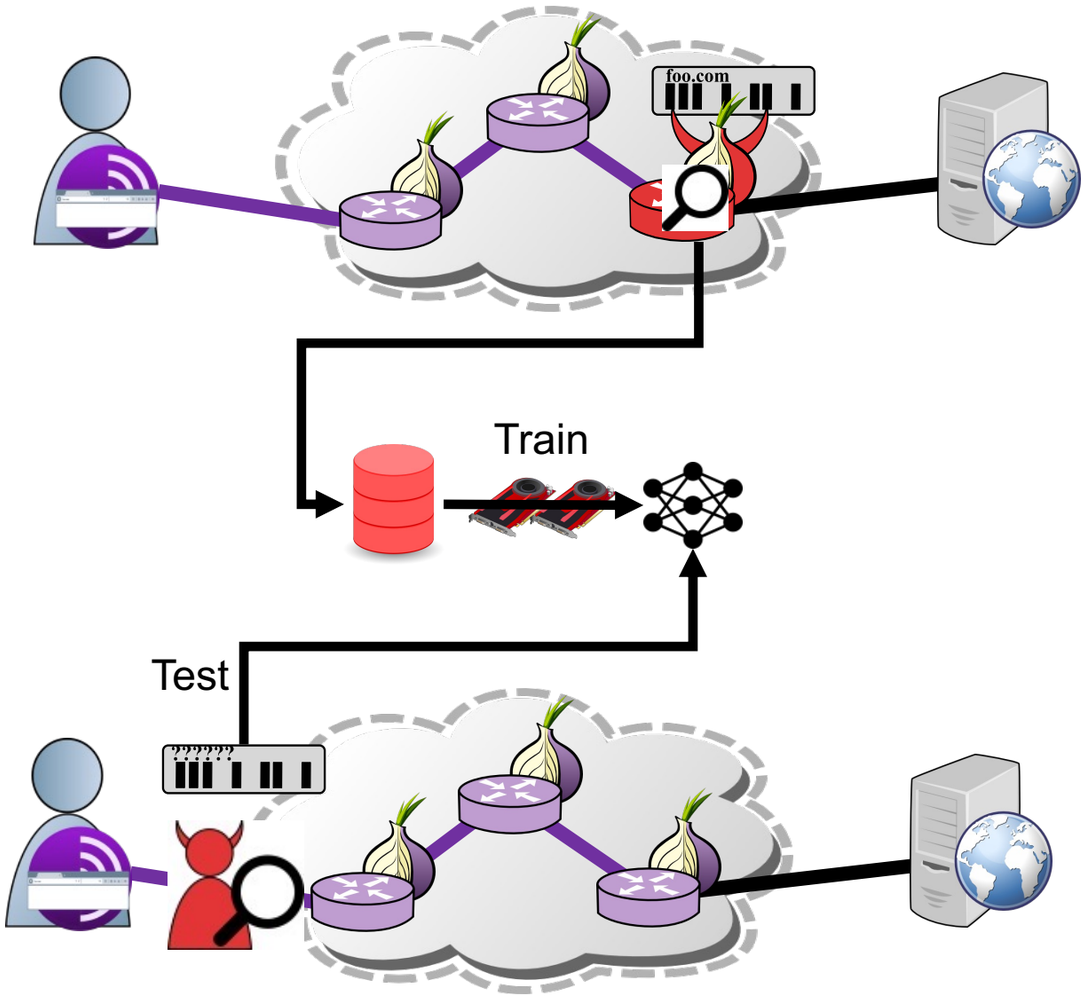
Real-World Evaluation Goals

We consider an adversary that uses real-world traces

- Real: traces from normal Tor users
- Testing ***must*** be against real traces
- Training on real traces is thus superior

We want to estimate WF performance as realistically as possible

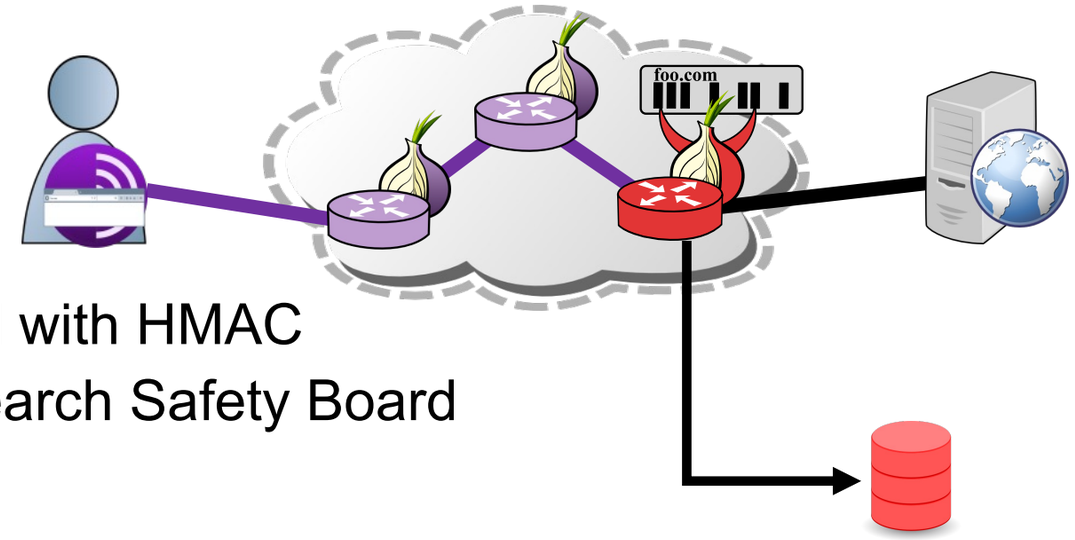
- Considering multiple training strategies
- We need a source of real-world data!



U.S. NAVAL RESEARCH LABORATORY

A Measurement of Genuine Tor Traces

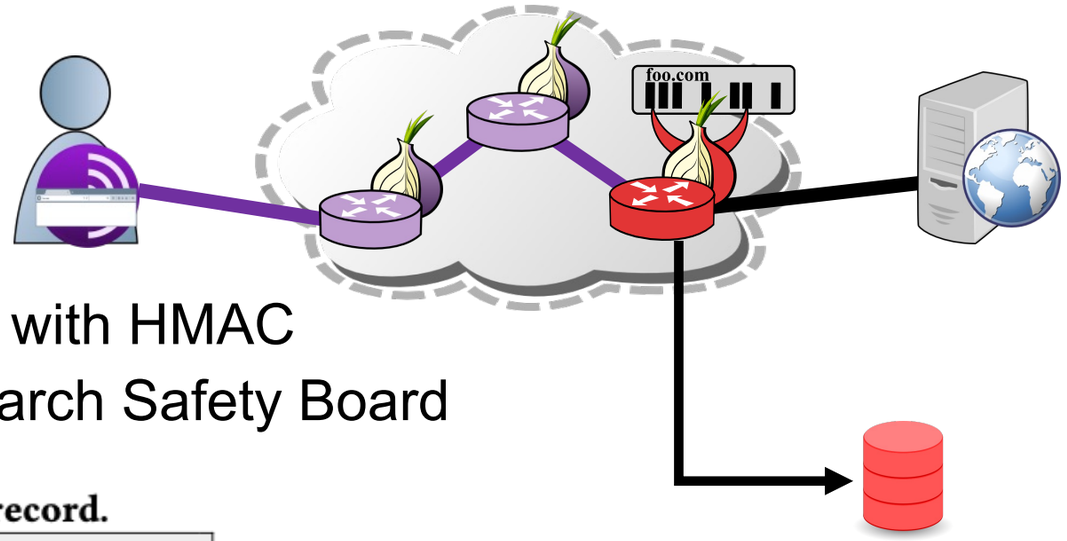
- Measurement from Tor exit relays
  - Measure normal Tor users at natural base rates
  - Sampling to limit data volume
  - No PII is recorded, metadata is protected with HMAC
  - Measurement plan reviewed by Tor Research Safety Board



GTT23
Genuine Tor Traces

A Measurement of Genuine Tor Traces

- Measurement from Tor exit relays
  - Measure normal Tor users at natural base rates
  - Sampling to limit data volume
  - No PII is recorded, metadata is protected with HMAC
  - Measurement plan reviewed by Tor Research Safety Board



GTT23
Genuine Tor Traces

Listing 1: Example circuit metadata record.

```
{
  "day":2,
  "domain":Dnqty37vYTIEivWhAEikb7HlJOzWXEZ2Rw05iicG7e8,
  "shortest_private_suffix":
       bIKFK8gYicwptEMM1Goxlo7KredMMFx48VD0MpXn9zc,
  "port":443,
  "cells":[
    [ 0.000015, 1,10,0],//client->exit: create
    [ 0.000463,-1,11,0],//exit->client: created
    [10.932340, 1, 9,1],//client->exit: relay_early.begin
    [12.070954,-1, 3,3],//exit->client: relay.connected
    [13.421017, 1, 9,2],//client->exit: relay_early.data
    [13.421030,-1, 3,2],//exit->client: relay.data
  ]
}
```

U.S. NAVAL RESEARCH LABORATORY

GTT23 Characteristics 1

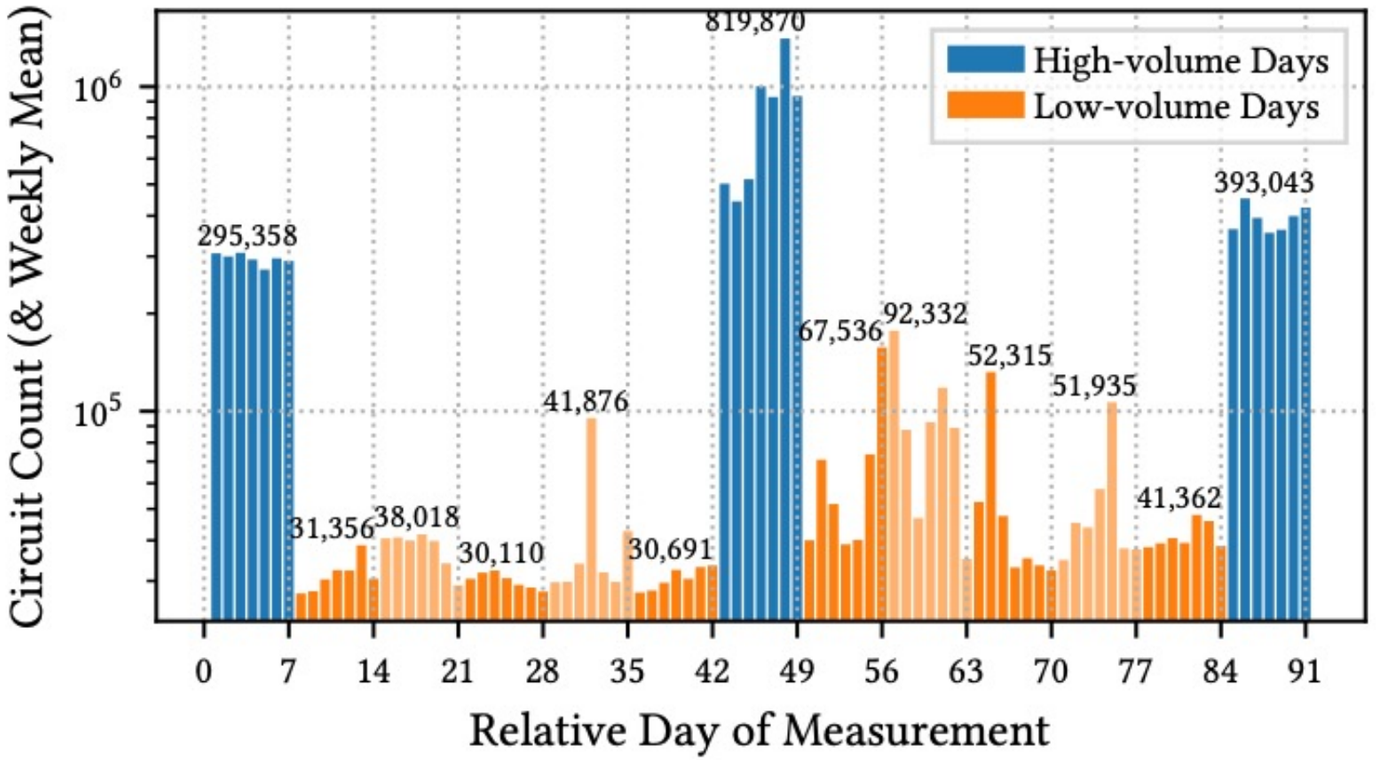


Figure 1: The daily total (bars) and weekly mean (text) number of circuits during our 13 week measurement.

GTT23 Characteristics 2

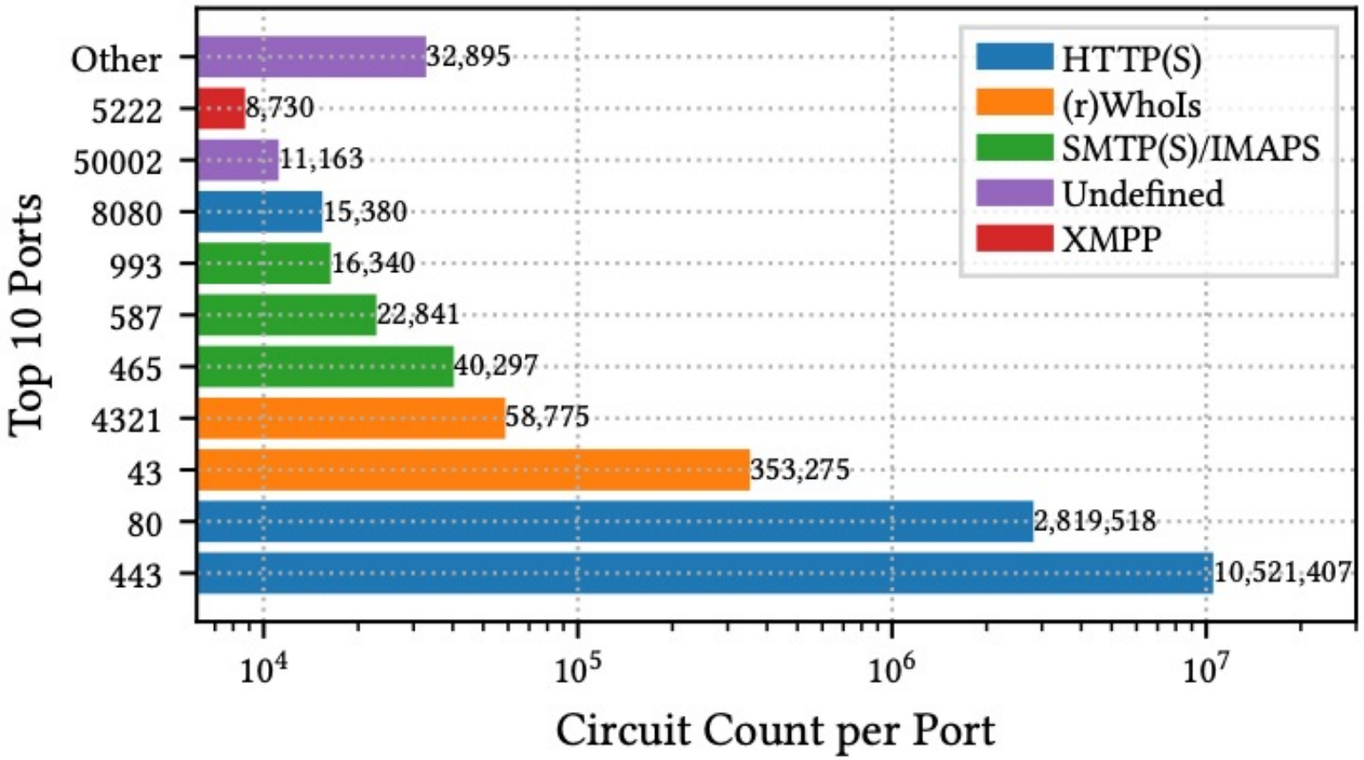


Figure 2: The total number of GTT23 circuits by server port, with IANA-assigned service names [45].

GTT23 Characteristics 3

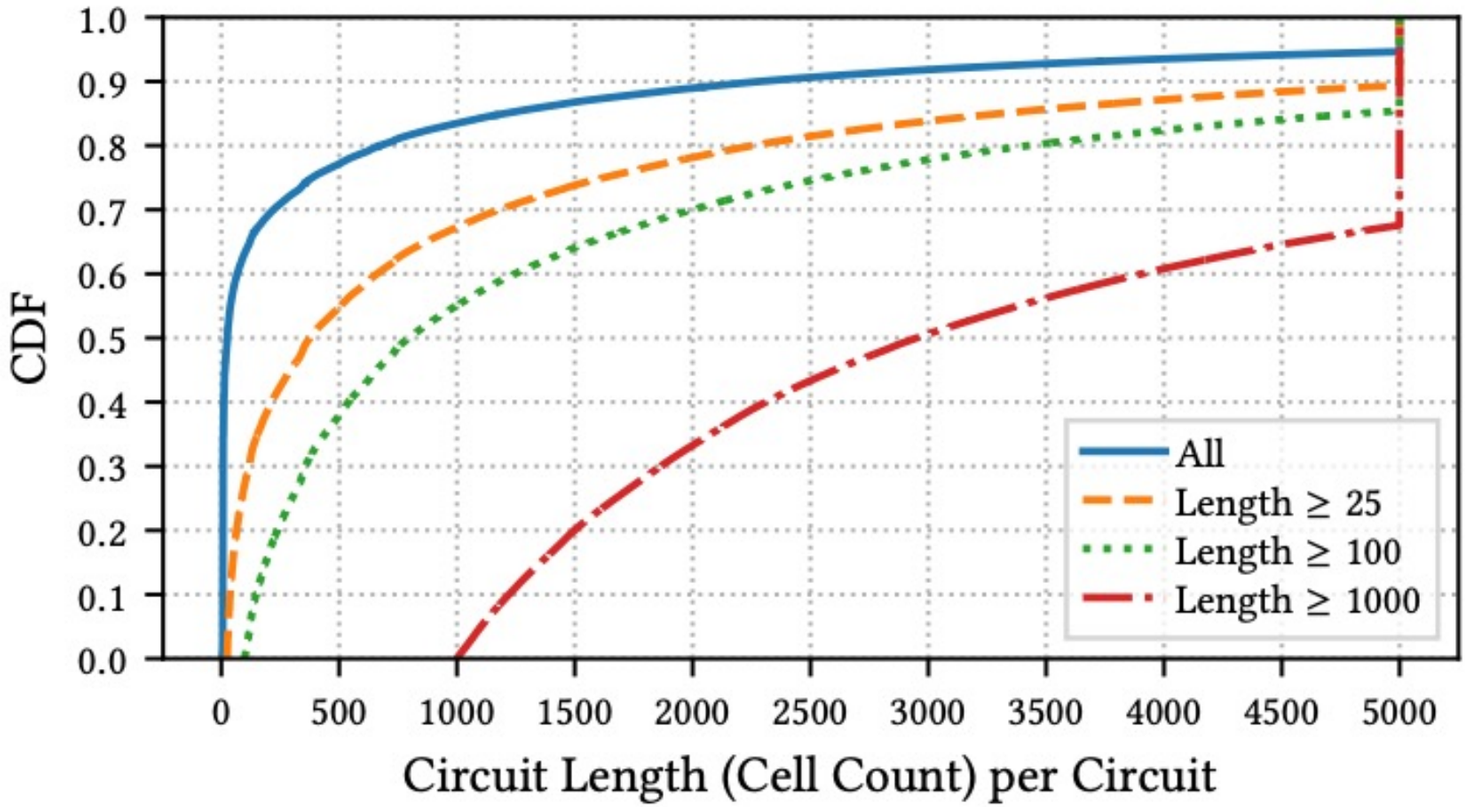


Figure 3: Cumulative distribution of the number of cells per circuit over subsets of GTT23 circuits.

GTT23 Characteristics 4

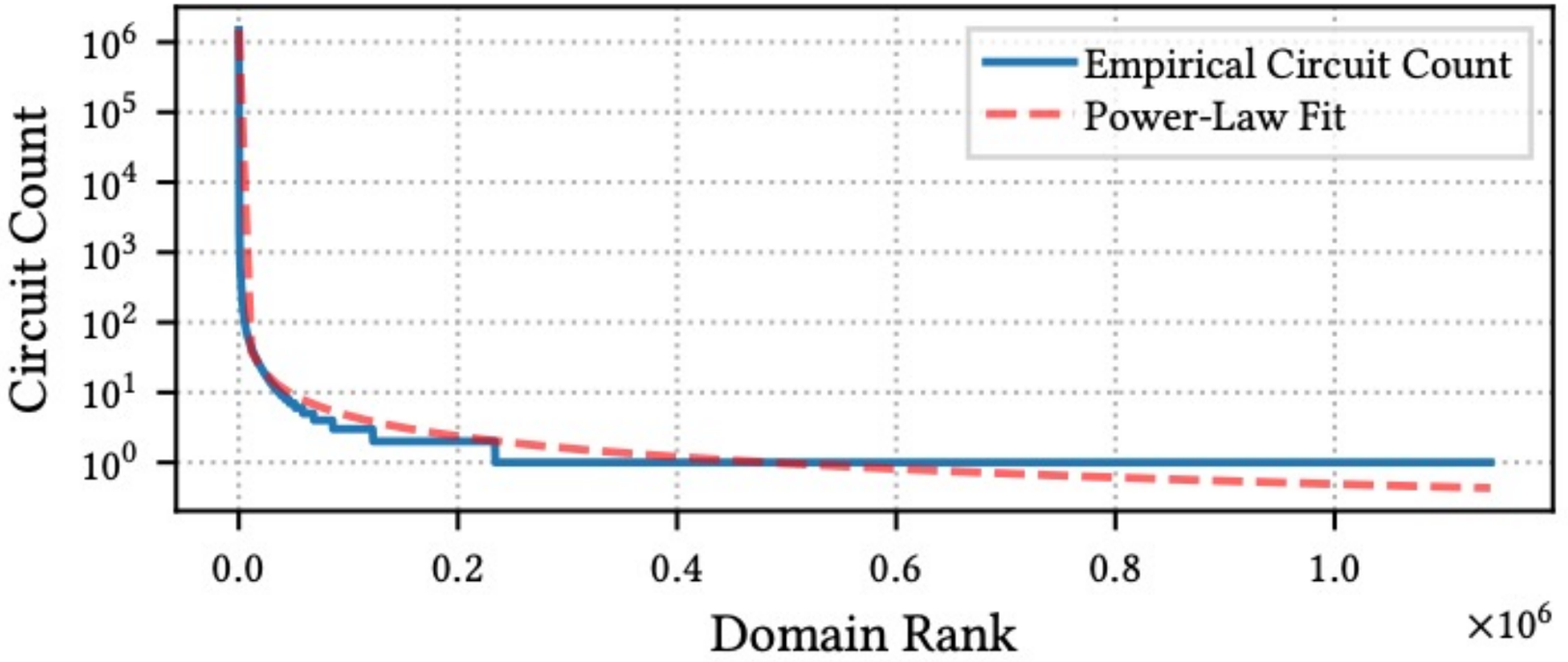


Figure 4: The number of GTT23 circuits per domain; we observe a close fit to a power-law distribution.

U.S. NAVAL RESEARCH LABORATORY

Comparisons to Synthetic Datasets

Table 3: Summary of website fingerprinting datasets curated over the past 15 years. The '⊥' symbol is used to indicate a dataset is unnamed, and the '-' symbol is used when a cell's contents are identical to the above cell. When the year of data collection is not mentioned, we assume it is around ("ca.") the associated article's publication date. Not all datasets describe their trace generation software with the same specificity. N, N_C, N_I, N_{Bg} are the total number of traces in the dataset, the number of positive classes, the number of instances per positive class, and the number of background traces. The "Attacks" column shows a list of WF attack papers evaluated on the dataset.

Ref.	Name	Year	Activity	Activity Detailed	User Model	Trace Gen. Software	N	N_C	N_I	N_{Bg}	Available	Attacks
[15]	⊥ (Hermann)	2008	Web	Links from real-world academic proxy server	Index page	Autofox	8.5×10^3	775	≈10		Dead link	[15]
[9]	⊥ (Cai)	Ca. 2012	Web	Alexa top sites	Index page	tor 0.2.1/2	3.2×10^4	800	≈40		No	[9]
[48]	levdata2	Ca. 2013	Web	Alexa top sites	Index page	tor 0.2.4.7; TBB 2.4.7	4×10^3	100	40		Online	[30, 48]
-	levdata3	-	-	Popular blocked sites, Alexa top sites	-	-	9×10^2	4	10	8.6×10^2	-	-
[47]	k-NN	Ca. 2014	Web	Sensitive sites, Alexa top sites	Index page	TBB 3.5.1; iMacros 8.6.0	1.4×10^4	100	90	5×10^3	Online	[1, 29, 30, 39, 47–49]
[23]	⊥ (Juárez)	Ca. 2014	Web	Alexa top sites, volunteer browsing	Index page, visited pages	TBB (2/3.X); Selenium	4.3×10^4	200	≈40	3.5×10^4	On request	[23]
[49]	⊥ (Wang)	2014	Web	Sensitive sites, Alexa top sites	Index page	tor 0.3.6.4; TBB 3.6.4	9×10^3	100	40	5×10^3	No	[49]
[30]	RND-WWW	Ca. 2016	Web	Twitter, Alexa one-click, Google Trends, Google Random, censored sites	Random subpage	TBB 3.6.1; Chickenfoot; iMacros; Scriptish	1.6×10^5	1,125	40	1.2×10^5	Dead link	[30]
-	TOR-Exit	-	-	HTTP requests of real Tor users	Visited page	-	2.1×10^5			2.1×10^5	-	-
-	WEBSITES	-	-	Popular websites	Index page, random subpage	-	5.3×10^3	50	105		-	-
[14]	DS_{Tor}	Ca. 2016	Web	Alexa top sites, popular .onion sites	Index page	TBB; Selenium	1.1×10^5	85	≈90	1×10^5	Dead link	[14, 29]
[36]	AWF CW_{900}	2017	Web	Alexa top sites	Index page	tor 0.2.8.11; TBB 6.5; Selenium	2.3×10^6	900	2,500		Online	[5, 28, 29, 36, 39]
-	AWF Recollect	-	-	-	-	-	1×10^5	200	500		-	-
-	AWF Open	-	-	-	-	-	8×10^5	200	2,000	4×10^5	-	-
[38]	DF	Ca. 2018	Web	Alexa top sites	Index page	tor-browser-selenium	1.4×10^5	95	1,000	4.1×10^4	Online	[28, 35, 38, 39]
[29]	WTT-time	2018	Web	Alexa top sites	Index page	tor 0.4.0.8; tor-browser-crawler	8×10^4	100	300	5×10^4	On request	[29]
[33]	Good Enough	2020	Web	Alexa top pages, random subpage	Index page	TBB 9.0.2	2×10^4	500	20	1×10^4	Online	
[46]	⊥ (Wang)	2019	Web	Alexa top sites	Index page	tor 0.4.0.1; TBB 8.5a7	1×10^5	100	200	8×10^4	Partially Online	[46]
-	Wikipedia	-	-	Wikipedia browsing	Random subpage	-	2×10^4	100	100	1×10^4	-	-
[28]	GDLF-25	Ca. 2021	Web	Alexa top sites	Random subpage	tor-browser-crawler	9.4×10^4	2,400	39		On request	[28]
-	GDLF-OW	-	-	Links from Rimmer et al. [36]	Random subpage	-	7×10^4			7×10^4	-	-
[27]	BigEnough	2021	Web	Open PageRank top pages	Index page	TBB	3.8×10^4	950	20	1.9×10^4	On request	
[11]	Multi-tab	2022	Web	Alexa top pages	Index page (multi-tab)	TBB; Selenium	5.7×10^5				Online	[11]
[21]	D(tbs, tor)	2022	Web	Wikipedia browsing	Random subpage	tor-browser-selenium	2×10^4	98	200		Online	
[4]	Drift	Ca. 2023	Web	Popular websites, links from Rimmer et al. [36]	Index page	TBB 11.0.10; tor-browser-selenium 0.6.3	1.5×10^4	90	≈110	5×10^3	Online	[4]
	GTT23	2023	Any	Real Tor usage	Visited service	Real client software	1.4×10^7	⟨ 1.1×10^6 domains ⟩			On request	

U.S. NAVAL RESEARCH LABORATORY

Comparisons to Synthetic Datasets

- Most synthetic datasets are focused on website index (front) pages
- There was a trend at considering larger datasets, GTT23 is still larger
- No other dataset contains “genuine” traces naturally created from real Tor users

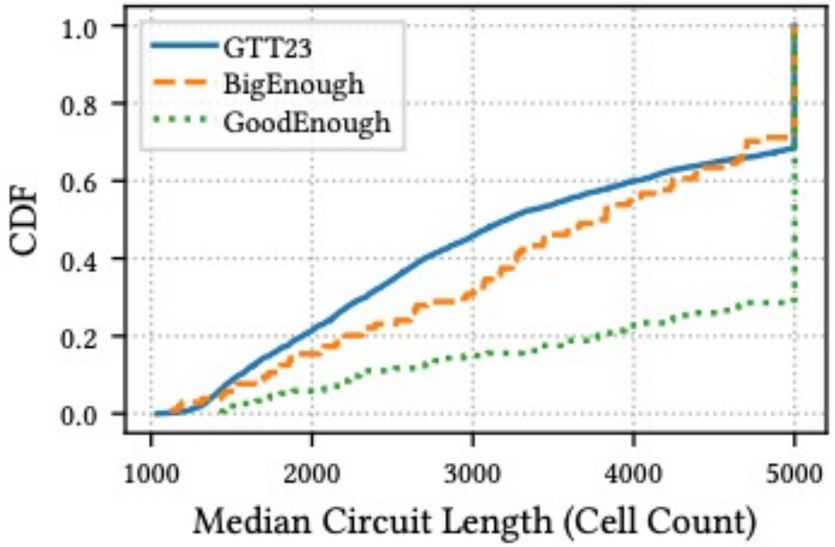
Table 2: Select WF Datasets (full details in Table 3)

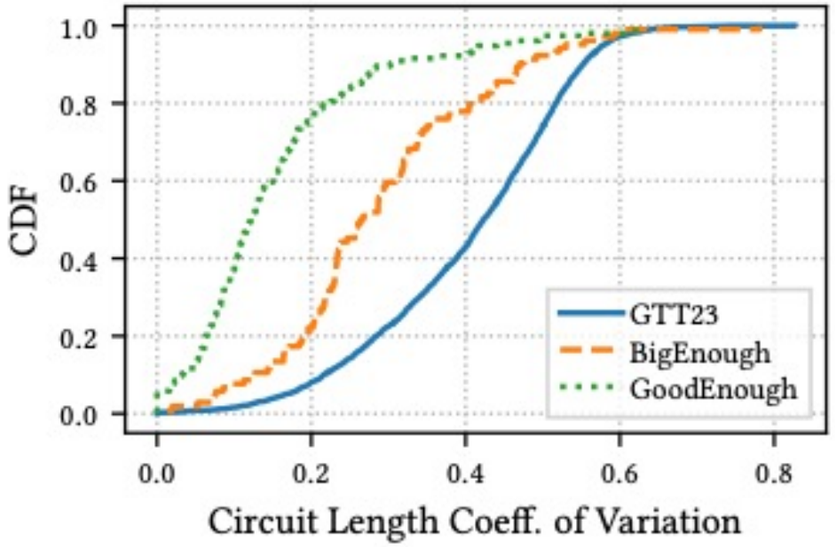
Dataset	Year	Size	Description†
k-NN [47]	2014	1.4×10^4	Web, top index pages
AWF CW_{900} [36]	2017	2.3×10^6	Web, top index pages
AWF Open [36]	2017	8×10^5	Web, top index pages
DF [38]	2018	1.4×10^5	Web, top index pages
GoodEnough [33]	2020	2×10^4	Web, top index pages + subpages
BigEnough [27]	2021	3.8×10^4	Web, top index pages + subpages
Multi-tab [11]	2022	5.7×10^5	Web, top index pages, multiple tabs
GTT23	2023	1.4×10^7	Genuine traffic, real user behavior, visited services, natural base rates

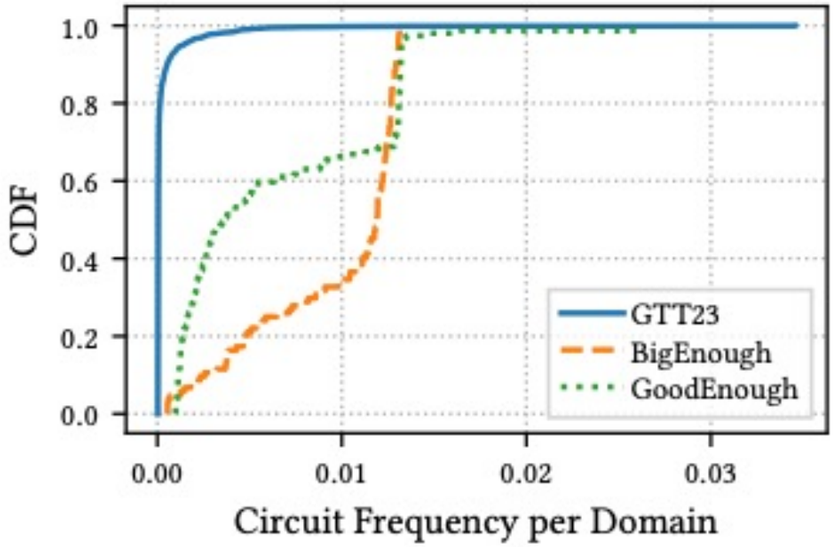
† All but GTT23 synthetically fetch webpages using automated tools.

U.S. NAVAL RESEARCH LABORATORY

Disparities between GTT23 and synthetic datasets







WF Methodology Considering Genuine Tor Traces

GTT23:

- Contains >13M traces from *real* users
- Collected over 13 weeks on Tor *exits*

GTT23 is available online:
Paper: https://doi.org/10.48550/arXiv.2404.07892
Dataset: https://doi.org/10.5281/zenodo.10620519

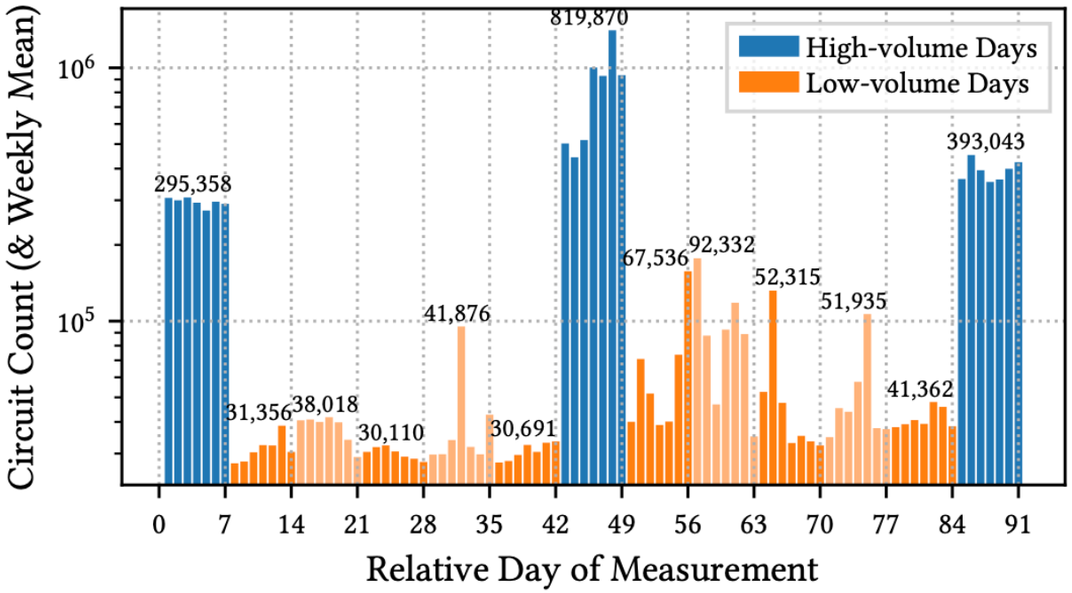


Figure 1: The daily total (bars) and weekly mean (text) number of circuits during our 13 week measurement.

Methodology Considering Genuine Tor Traces

GTT23:

- Contains >13M traces from *real* users
- Collected over 13 weeks on Tor *exits*

Training:

- Use Deep Fingerprinting (DF) model
- Week 1 traces with ≥ 1000 cells
- 1 model for each of the ~400 most popular websites

Testing

- Traces from weeks >1
- Open world: some sites not trained on

GTT23 is available online:
Paper: https://doi.org/10.48550/arXiv.2404.07892
Dataset: https://doi.org/10.5281/zenodo.10620519

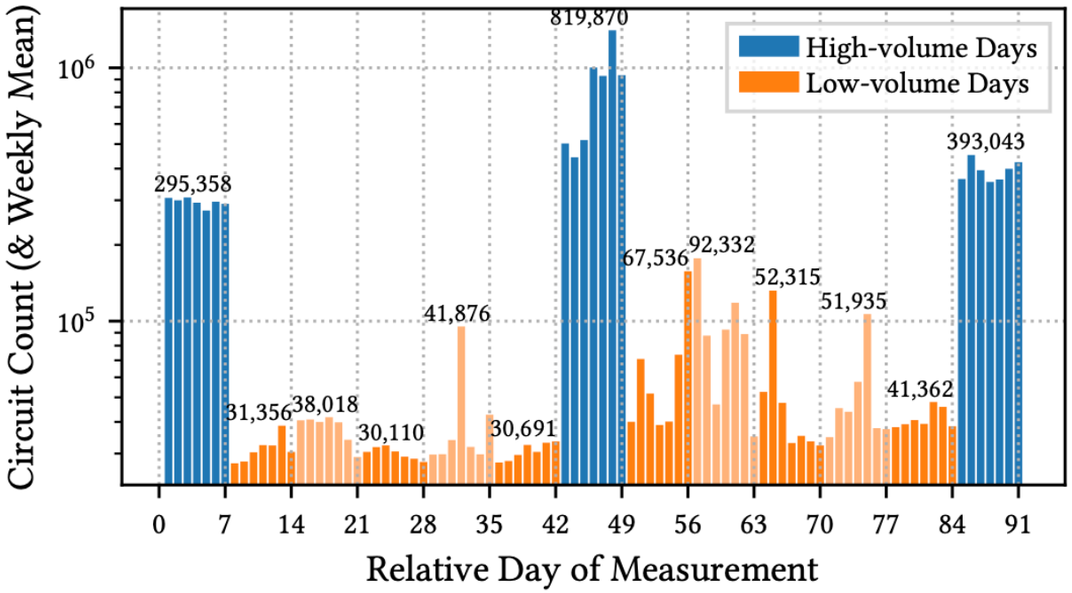
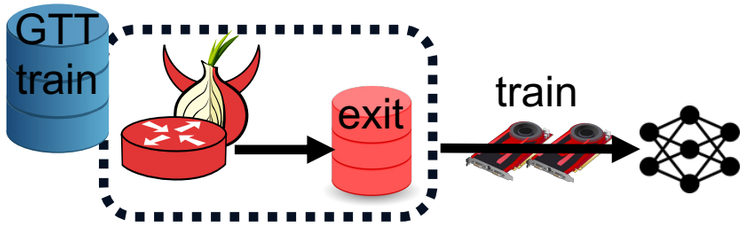


Figure 1: The daily total (bars) and weekly mean (text) number of circuits during our 13 week measurement.

WF Performance when Testing on Entry Traces

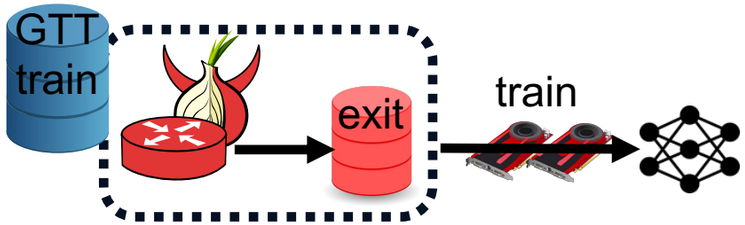
OnlineWF Train: (Cherubin'22)



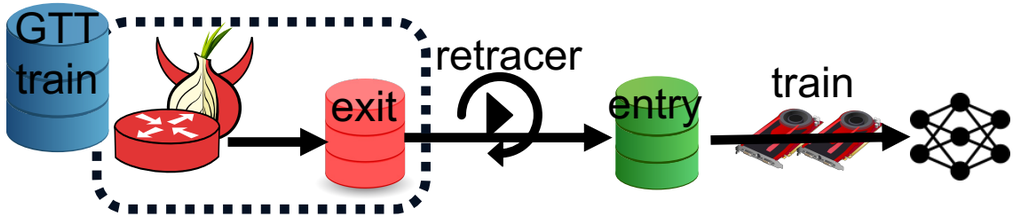
U.S. NAVAL RESEARCH LABORATORY

WF Performance when Testing on Entry Traces

OnlineWF Train: (Cherubin'22)



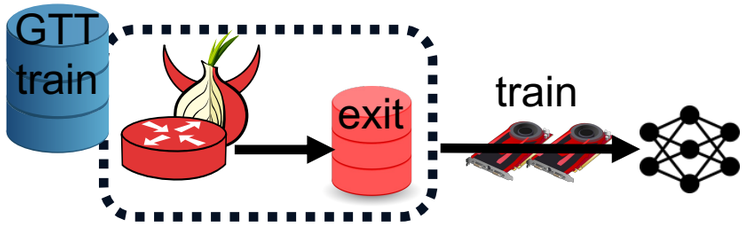
Retracer Train:



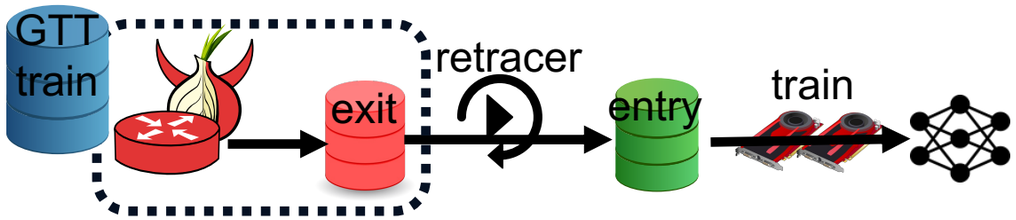
U.S. NAVAL RESEARCH LABORATORY

WF Performance when Testing on Entry Traces

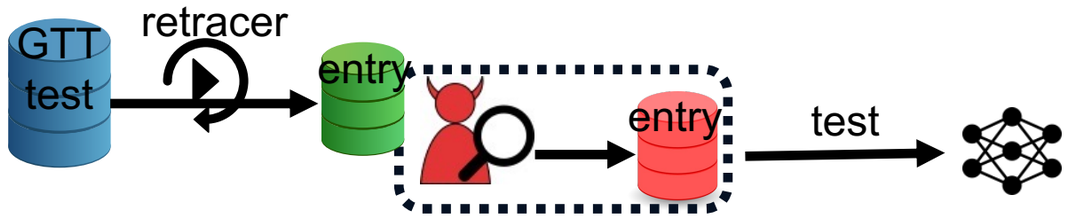
OnlineWF Train: (Cherubin'22)



Retracer Train:

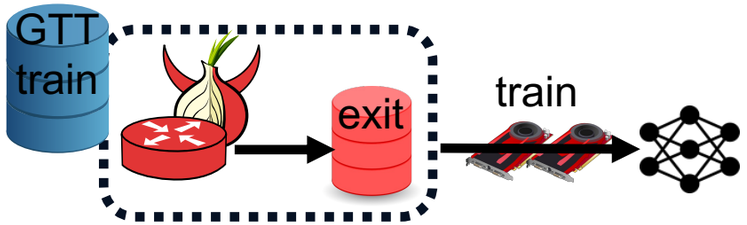


Both Test:

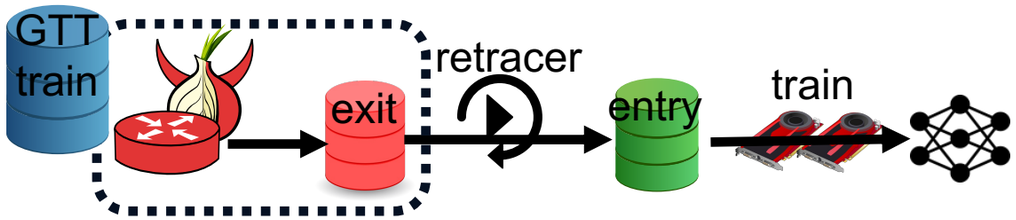


WF Performance when Testing on Entry Traces

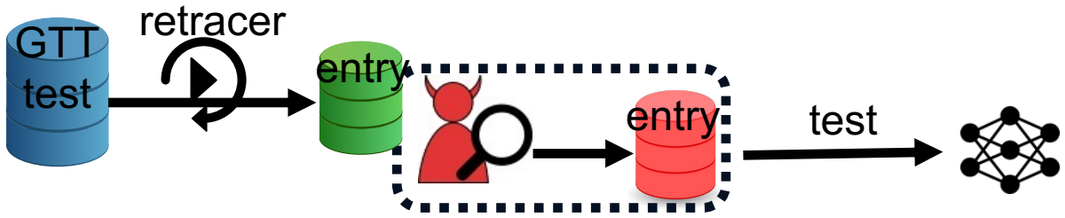
OnlineWF Train: (Cherubin'22)



Retracer Train:



Both Test:



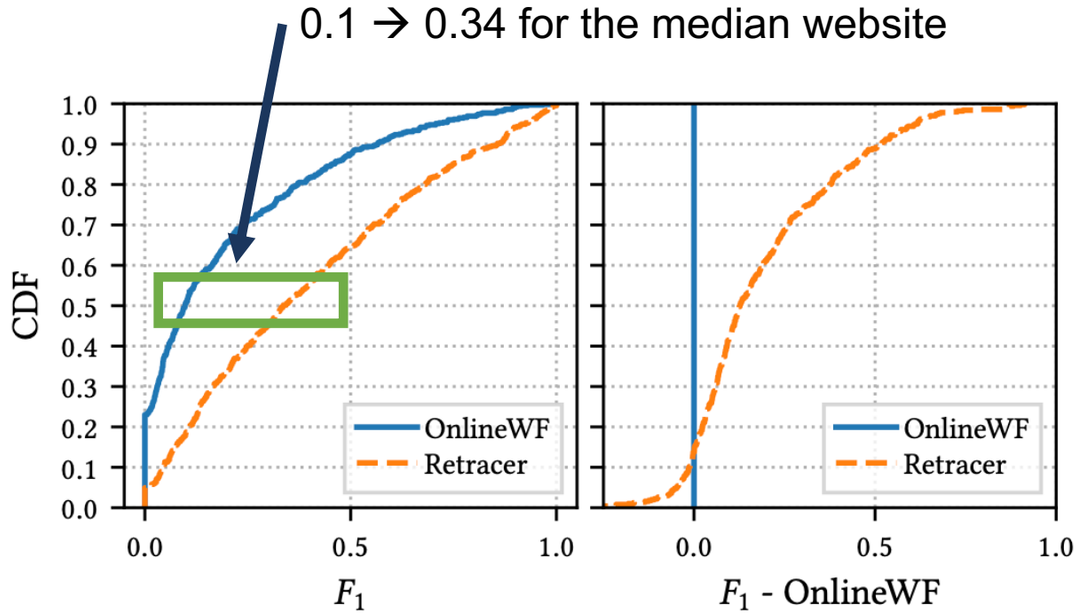


Figure 8: Classifier performance when training on exit traces as in OnlineWF [8] and training on entry traces transduced from the exit traces by Retracer.

U.S. NAVAL RESEARCH LABORATORY

Synthetic Datasets Overestimate WF Performance

Retracer: trained & tested as before

- Uses Retracer to transduce the GTT23 train and test sets

Synthetic datasets → previous work

- BigEnough: ~100,000 traces
- GoodEnough: ~10,000 traces
- Multiple pages per site
- Use analogous per-site training/testing methodology

Synthetic Datasets Overestimate WF Performance

Retracer: trained & tested as before

- Uses Retracer to transduce the GTT23 train and test sets

Synthetic datasets → previous work

- BigEnough: ~100,000 traces
- GoodEnough: ~10,000 traces
- Multiple pages per site
- Use analogous per-site training/testing methodology

WF performs better with synthetic traces

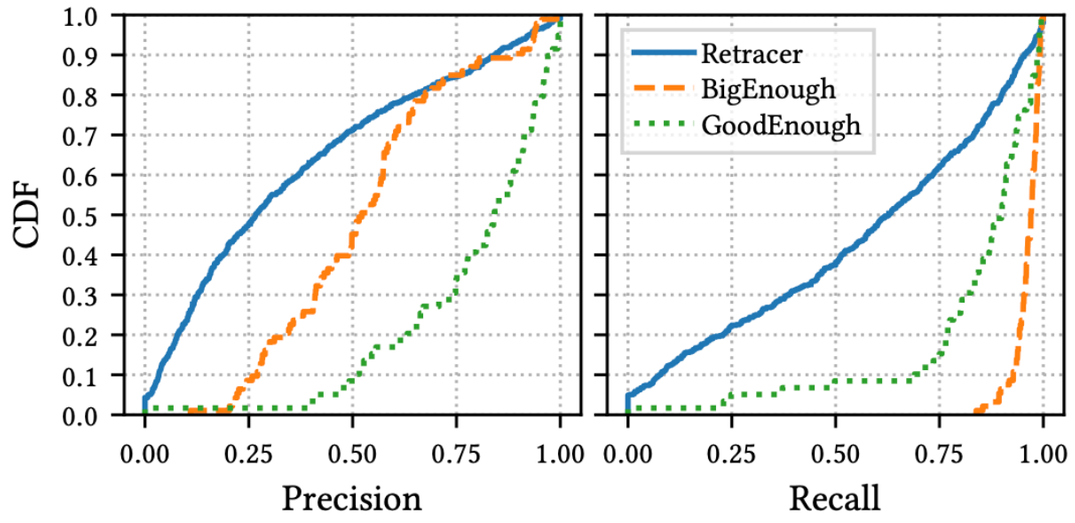


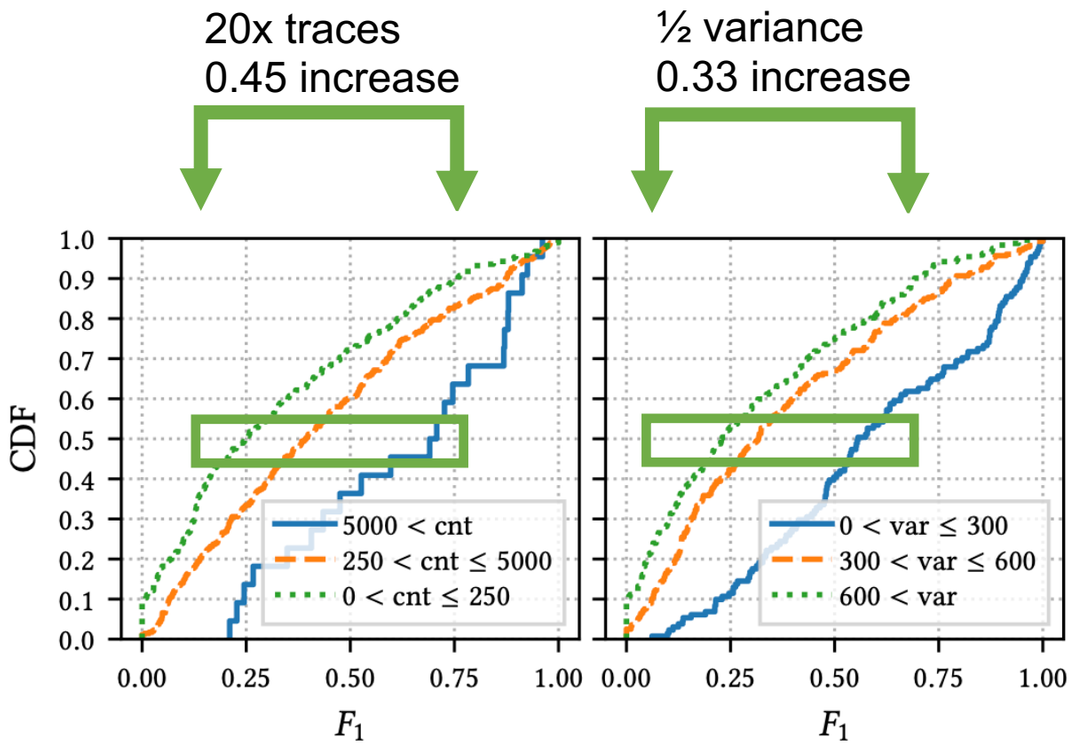
Figure 9: Performance of the classifiers trained and tested with each dataset. "Synthetic" traces lead to better performance than Retracer traces (transduced from GTT23).

U.S. NAVAL RESEARCH LABORATORY

What are the important features for performance?

Feature importance analysis – features predicting performance

1. Trace count
   - Median F_1 increased by 0.45 when 20x as many traces were available
2. Variance of trace lengths
   - Median F_1 increased by 0.33 when half as much variance is observed



U.S. NAVAL RESEARCH LABORATORY

Repositioning Real-World Website Fingerprinting on Tor

Contributions

- Retracer for transducing cell traces across positions
- Retracer evaluation using Tor datasets
- Real-world WF evaluation that tests on entry traces
- Individual website fingerprintability methodology
- Feature importance analysis

Future Work

- Use Retracer to evaluate WF in new scenarios
  - Traffic splitting with Conflux
  - Apply WF defenses on top of genuine data
- New methods for transduction

Read the paper!

Contact:
robert.g.jansen7.civ@us.navy.mil
robgjansen.com