Automated Website Fingerprinting through Deep Learning

Vera Rimmer[1], Davy Preuveneers[1], Marc Juarez[2],
Tom Van Goethem[1] and Wouter Joosen[1]

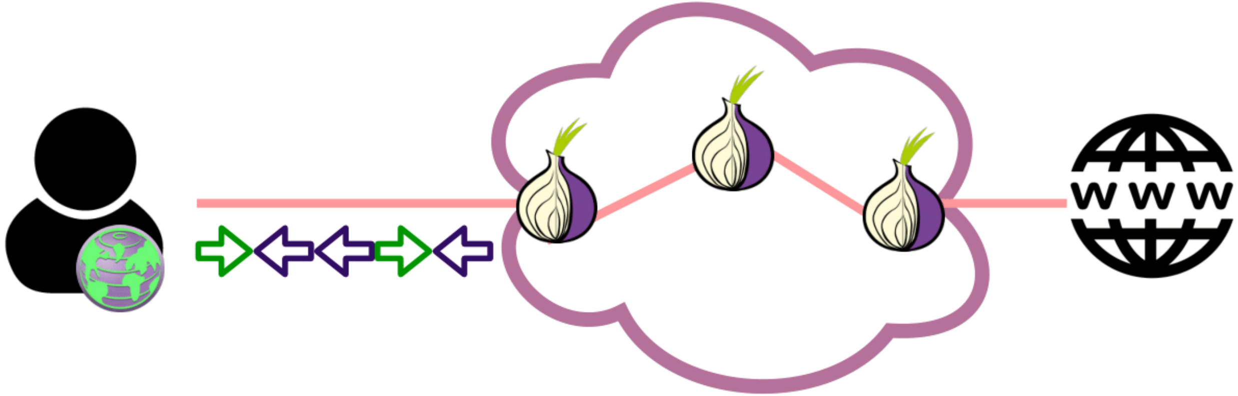
NDSS 2018 – Feb 19th (San Diego, USA)

DistriNet [1]
COSIC [2]
KU LEUVEN

Website Fingerprinting

Anonymous Communication through Tor

- All (secure) communication protocols expose metadata
  - timing, size of packets, identities, locations, addresses, communication patterns –> reveal private information
- Anonymity tools relay traffic through protected communication channels
  - The Onion Router (Tor)

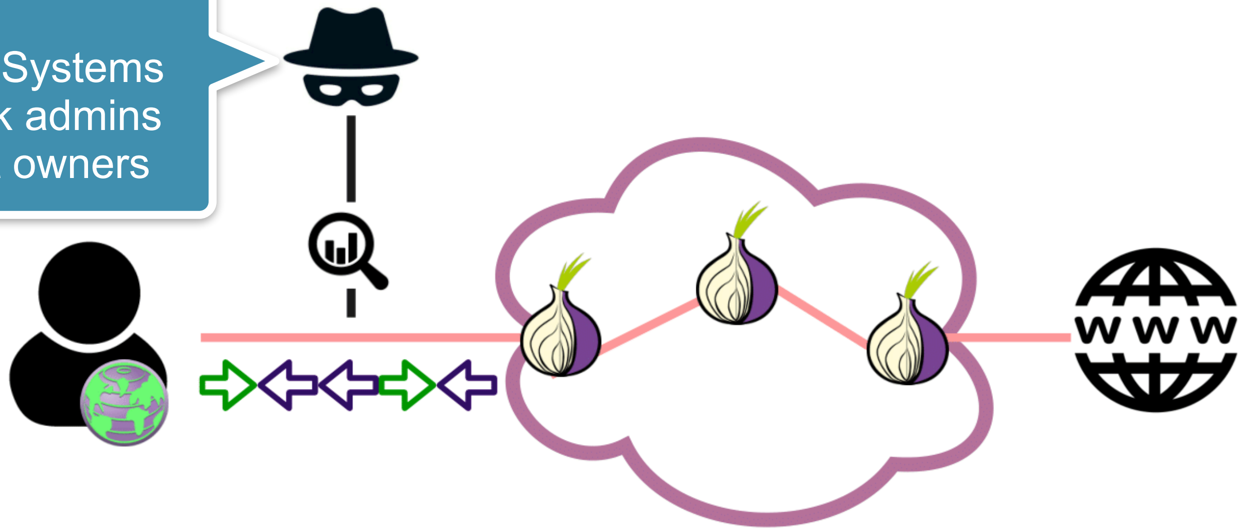


DistriNet

Website Fingerprinting

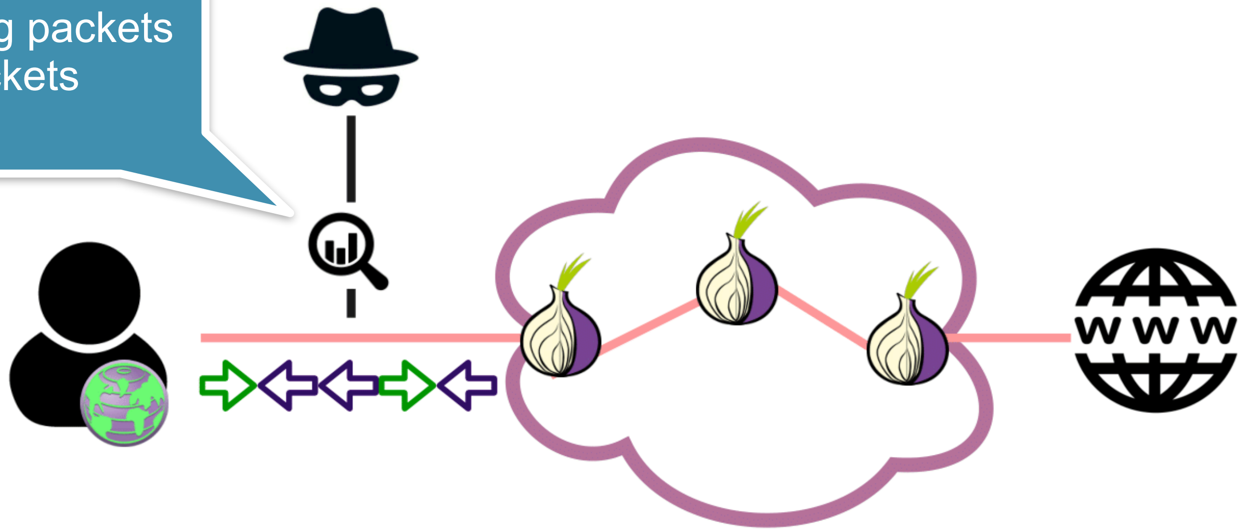
- Side-channel attack that reveals user's browsing activity
- Adversary is a local eavesdropper

- ISP
- Autonomous Systems
- Local network admins
- Wi-Fi hotspot owners

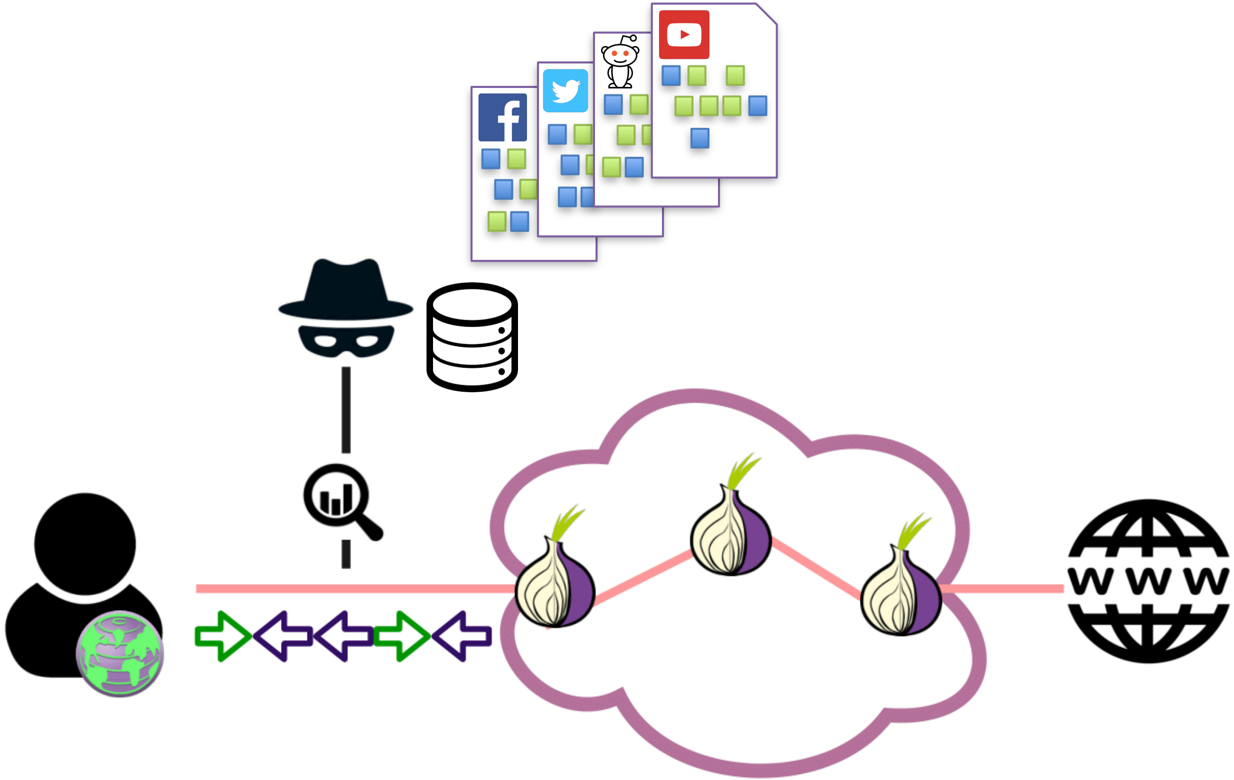


Website Fingerprinting

- Number of packets
- Average packet size
- % of incoming packets
- Timing of packets
- ...

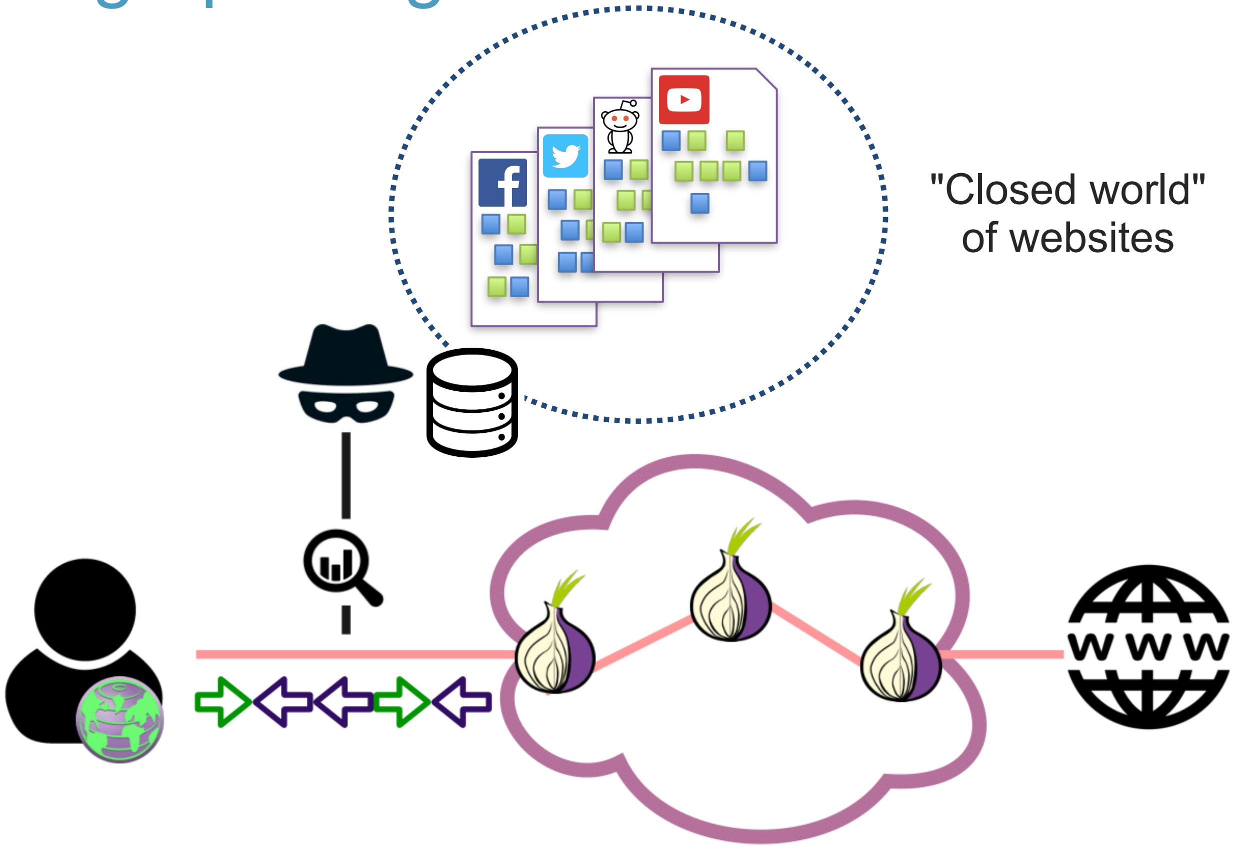


Website Fingerprinting

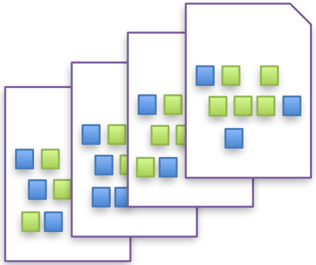


DistriNet

Website Fingerprinting



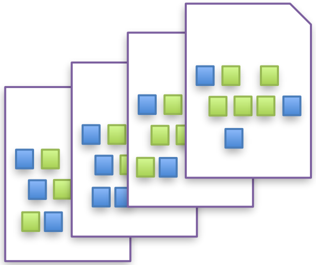
Website Fingerprinting Pipeline



Communication
patterns

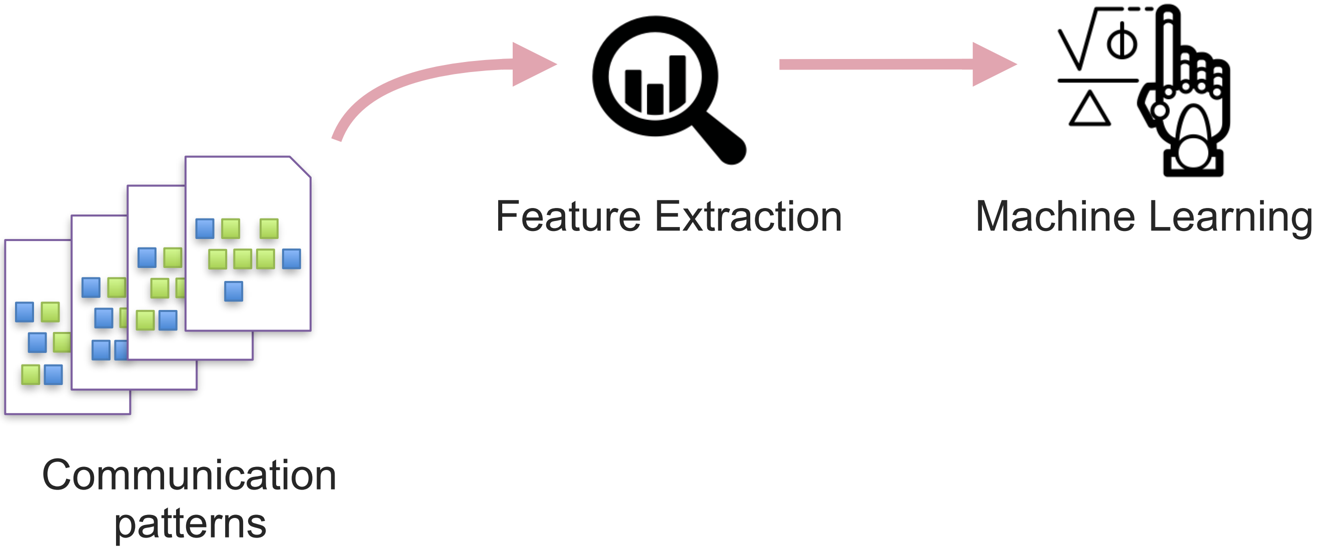
Website Fingerprinting Pipeline

Feature Extraction

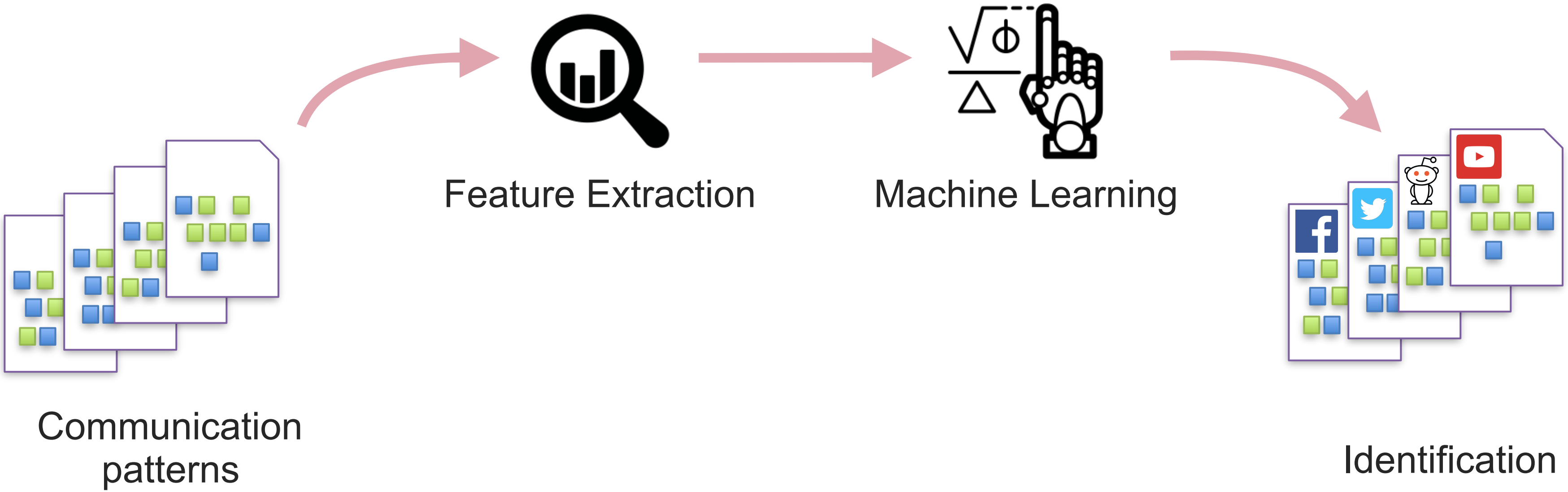


Communication
patterns

Website Fingerprinting Pipeline



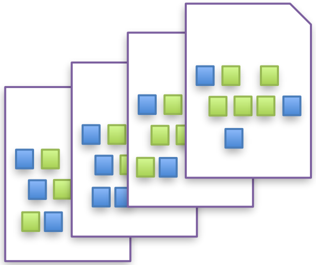
Website Fingerprinting Pipeline



State-of-the-Art Attacks

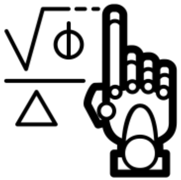
- kNN (Wang et al., 2014)
  - 3,000 features picked through heuristics (total size, total time, number of packets, packet ordering, traffic bursts…)
  - Classifier: k-Nearest Neighbors
- k-Fingerprinting (Hayes et al., 2016)
  - 150 features selected from Wang's through the analysis of feature importance
  - Classifier: Random Forest and k-Nearest Neighbors
- CUMUL (Panchenko et al., 2016)
  - 100 features, interpolation points of the cumulative sum of packet lengths
  - Classifier: Support Vector Machine

Website Fingerprinting Arms-race

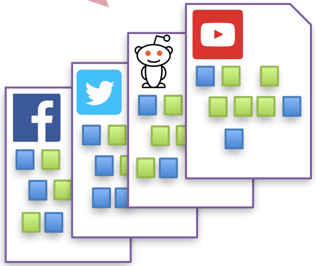


Communication patterns

Feature Extraction



Machine Learning



Identification

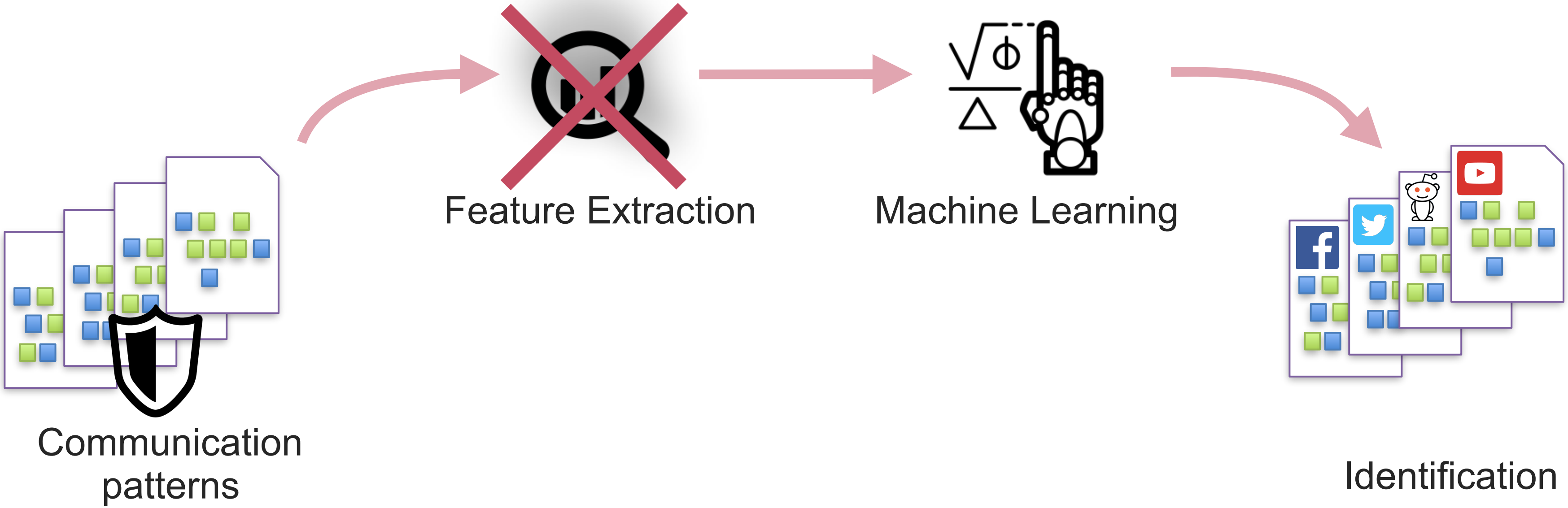
Main focus of prior work:
- Manual engineering
- Intellectual effort
- Difficult and expensive

AND

Success of attacks is defined by the set of engineered features

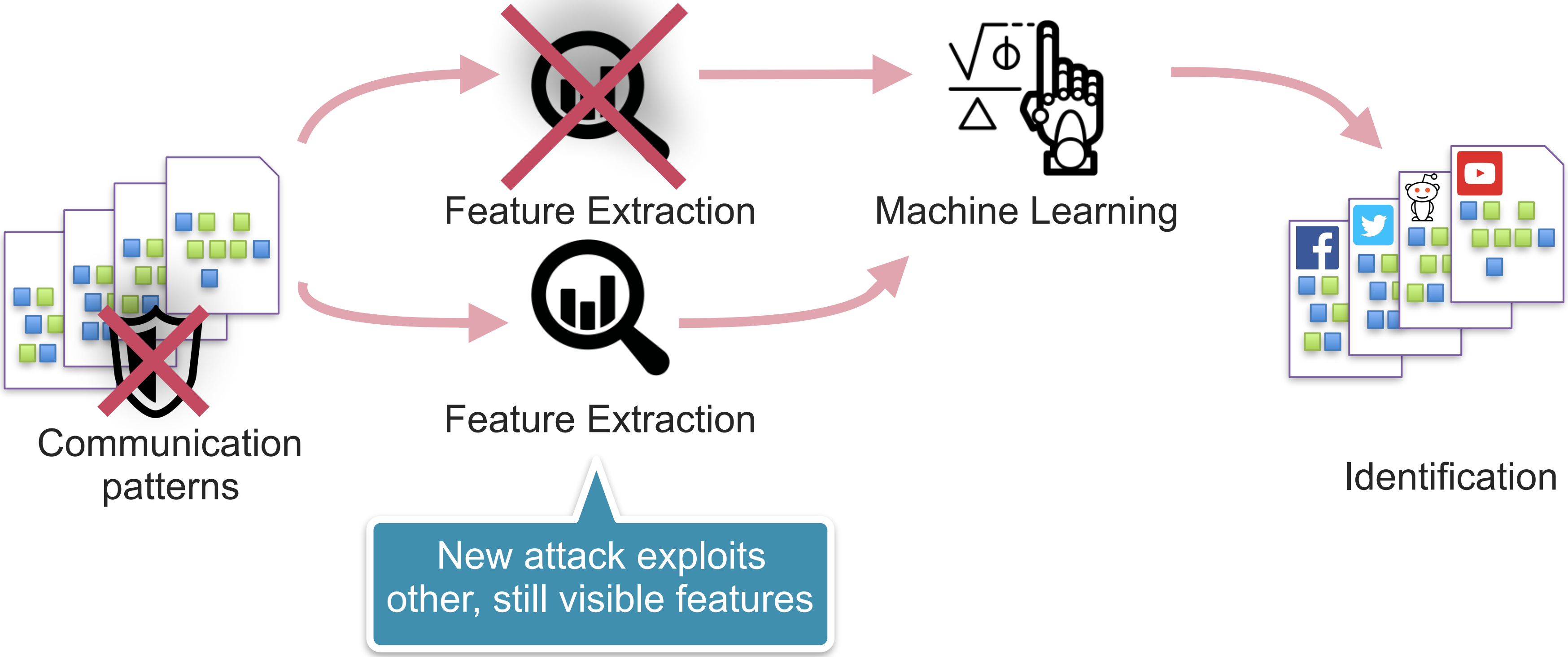
DistriNet

Website Fingerprinting Arms-race



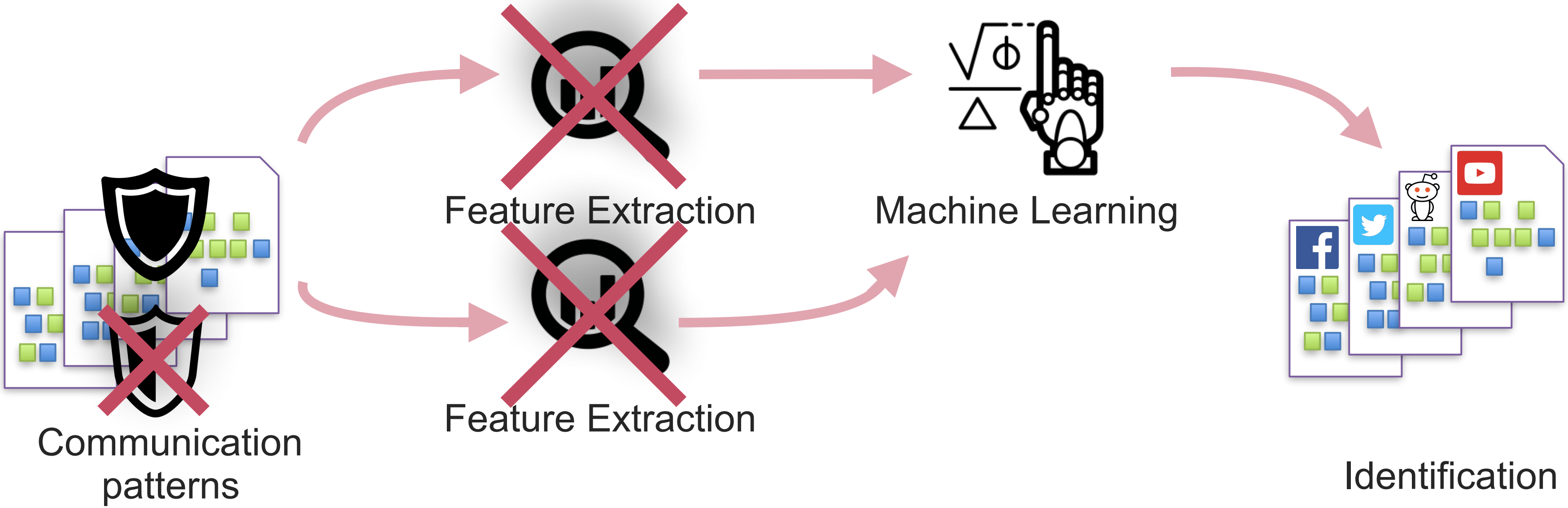
Concealing these features creates a countermeasure

Website Fingerprinting Arms-race



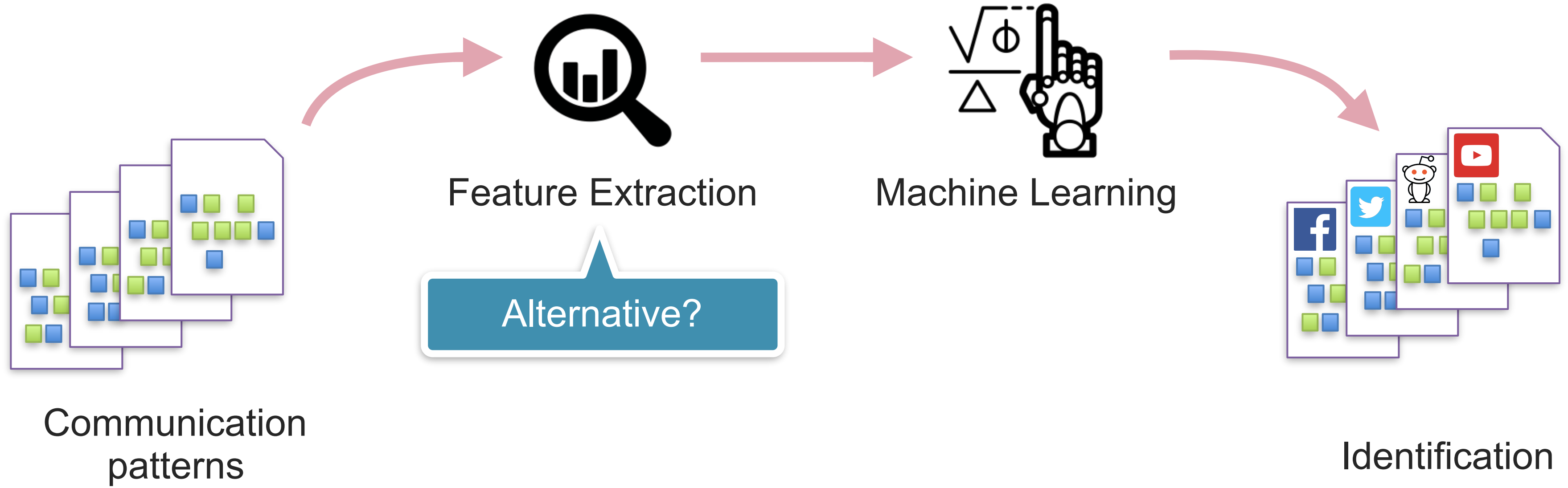
DistriNet

Website Fingerprinting Arms-race

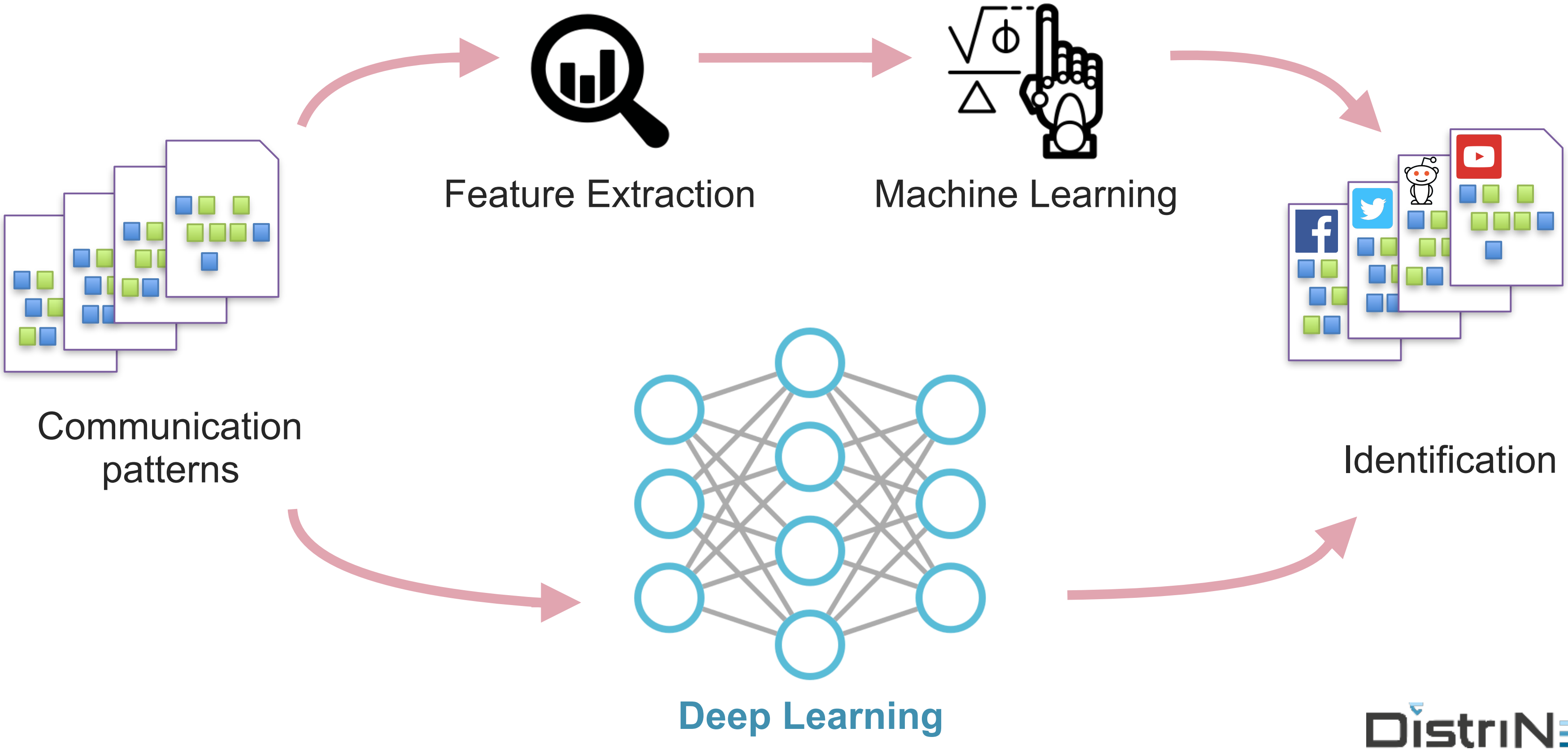


DistriNet

Website Fingerprinting



Website Fingerprinting



DistriNet

Deep Learning for WF

Why Deep Learning?

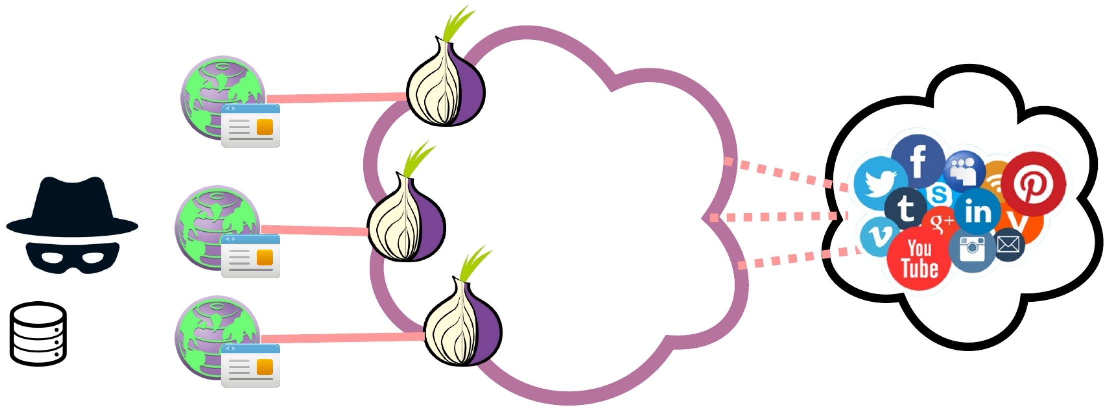
- Automatic feature learning from raw input
  - Obviates hand-engineering of features
  - Adaptive to changes in patterns
- Limited transparency and interpretability
  - Learned features are implicit and abstract
- Efficient, easily distributed and parallelized

Deep Learning based WF

- Data Collection
  - DL requires a lot of training data
- Deep Neural Network choice
  - Choosing the best suited deep learning algorithm
- Hyperparameter Tuning and Model Selection
  - Tuning of heavily parameterised models

Data Collection

- Built a distributed crawler
  - captures timing, direction and sizes of TCP packets
- 2,500 traces for each 900 top Alexa most popular sites: **largest-ever dataset**
- Closed worlds: CW_N datasets, where N is the number of sites



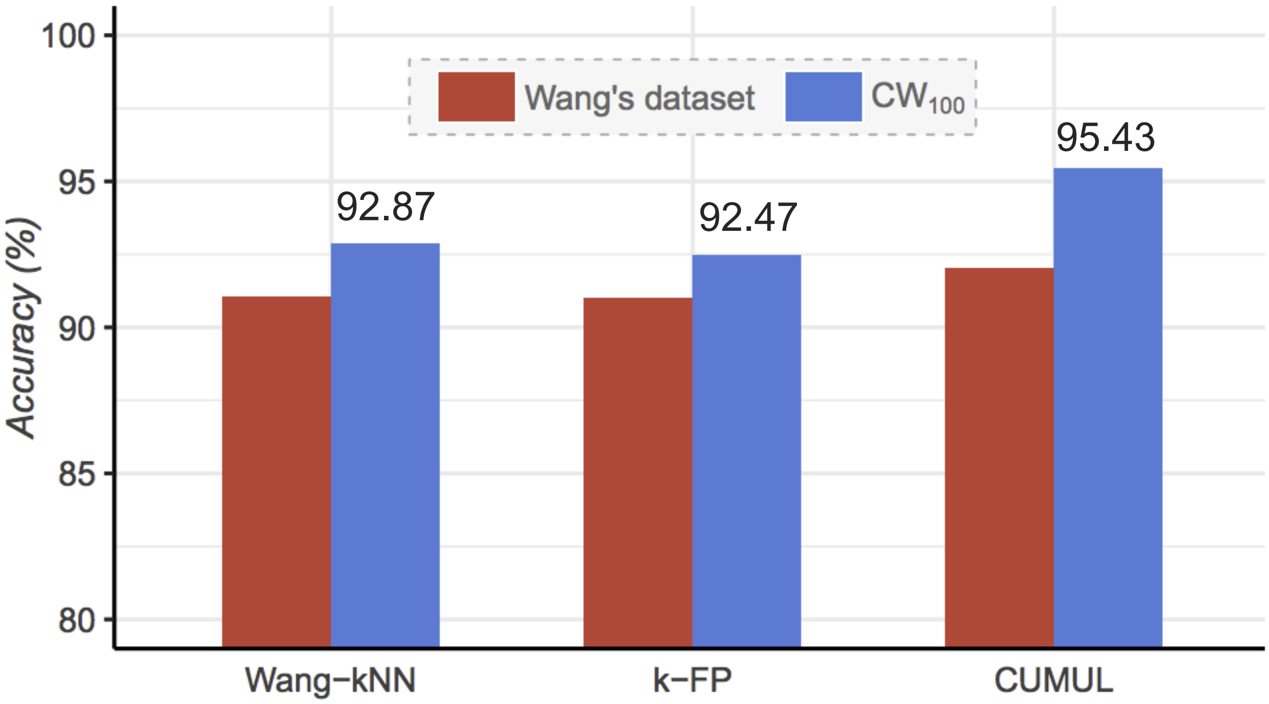
DistriNet

Deep Neural Networks

- Choice of a Deep Neural Network (DNN) suited for the input data
  - 1D sequences of incoming and outgoing Tor cells encoded as 1 and -1
- Explored 3 major types of DNNs:
  - feedforward: Stacked Denoising Autoencoder (SDAE)
    - learns from the *continuous structure* through dimensionality reduction
  - convolutional: Convolutional Neural Network (CNN)
    - learns from the *spatial structure* through convolutions and subsampling
  - recurrent: Long Short Term Memory (LSTM)
    - learns from the *temporal structure* (time-series) through internal memory

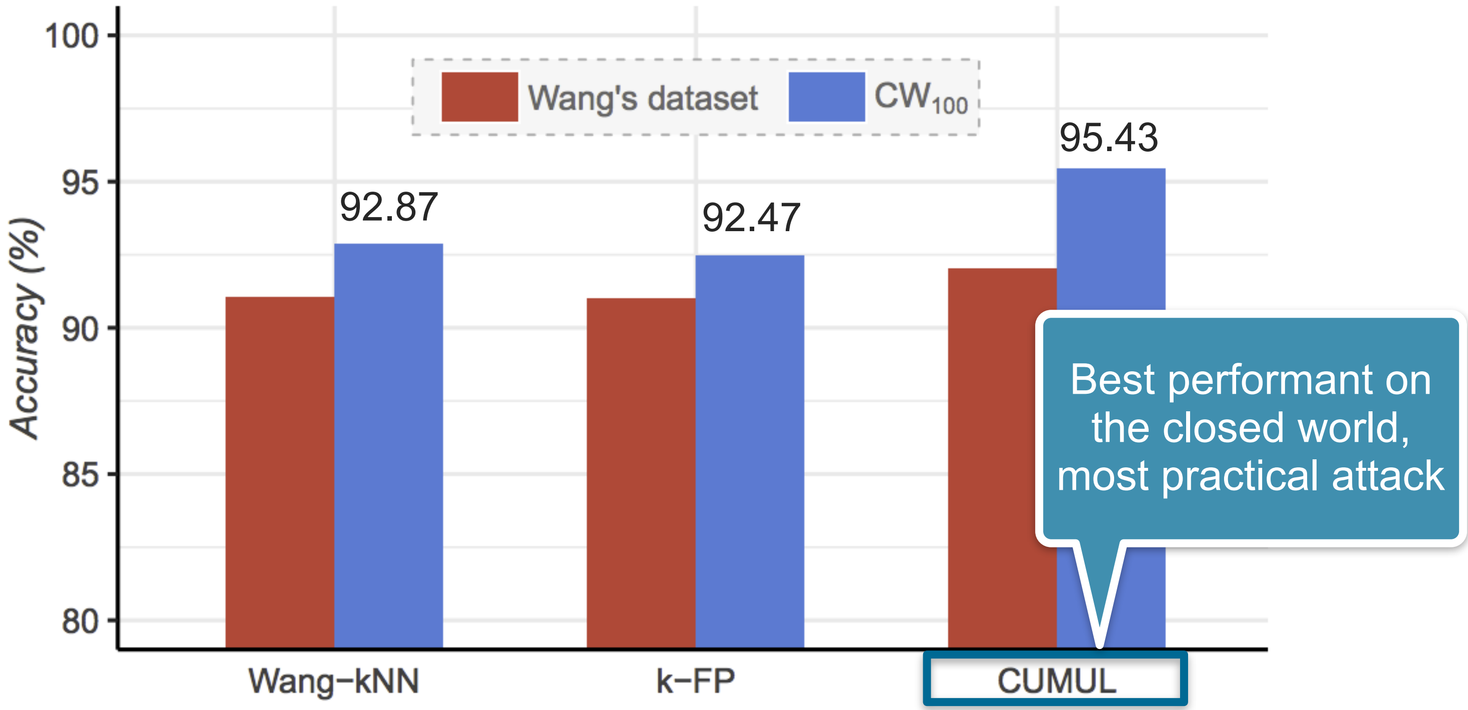
Evaluation and Results

Re-evaluation of Traditional Attacks



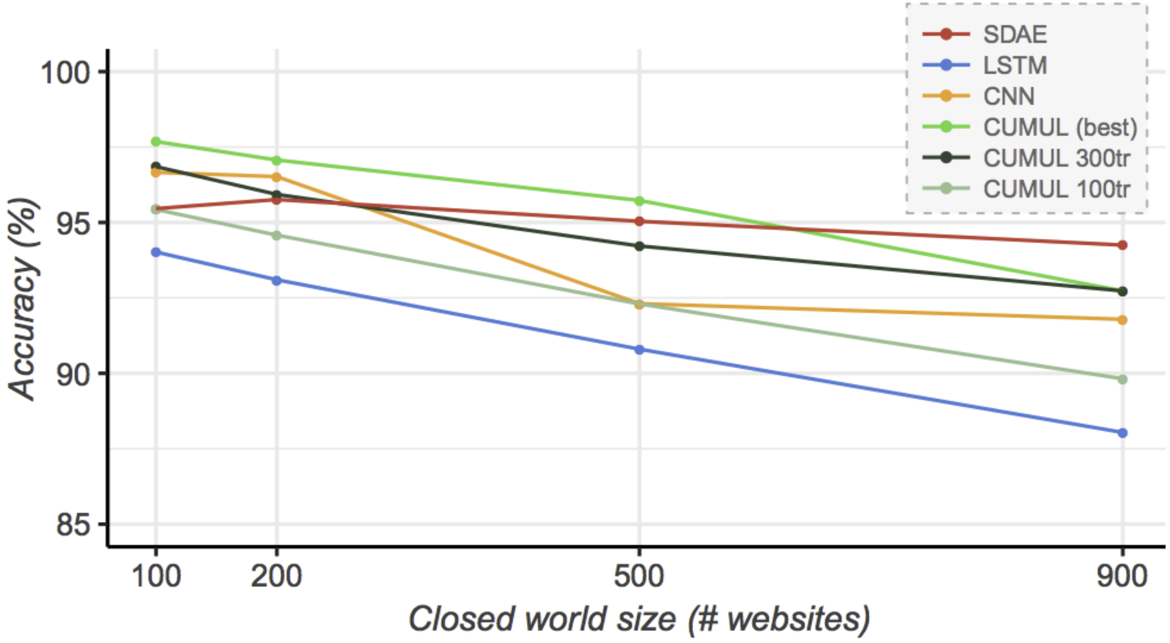
DistriNet

Re-evaluation of Traditional Attacks



DistriNet

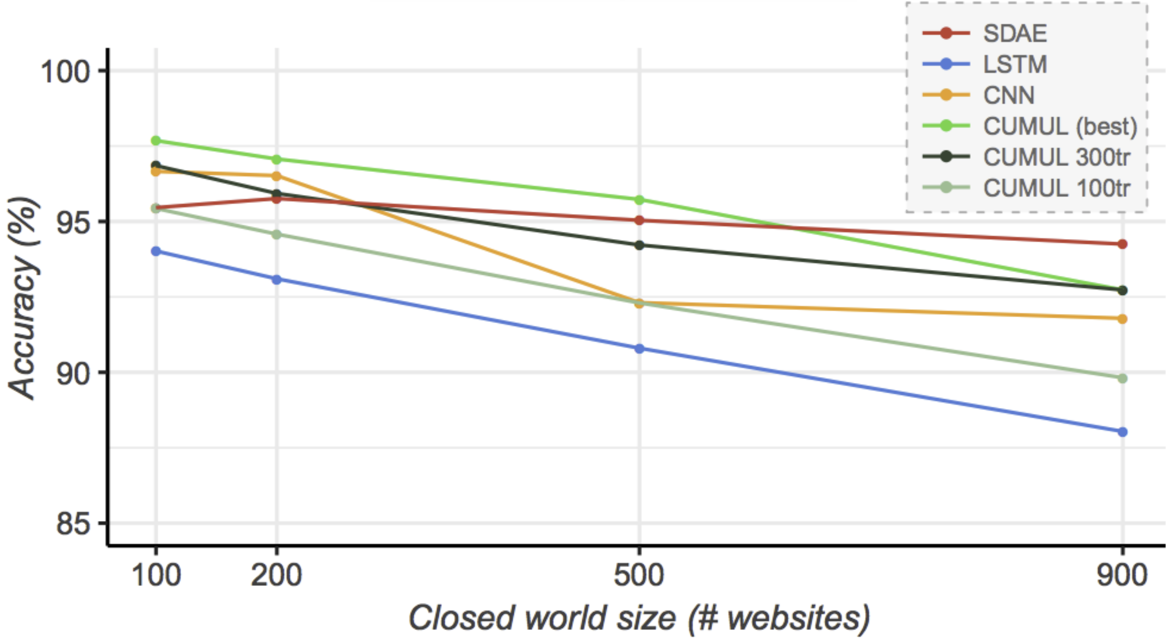
Closed World



DistriNet

Closed World

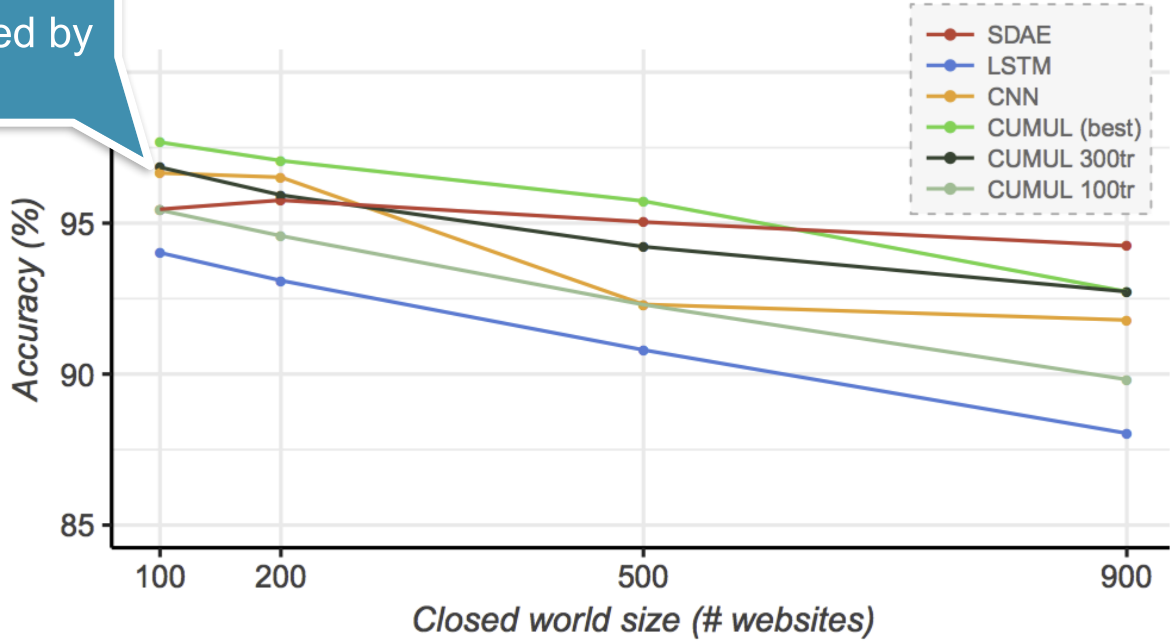
Overall, comparable with the state-of-the-art



DistriNet

Closed World

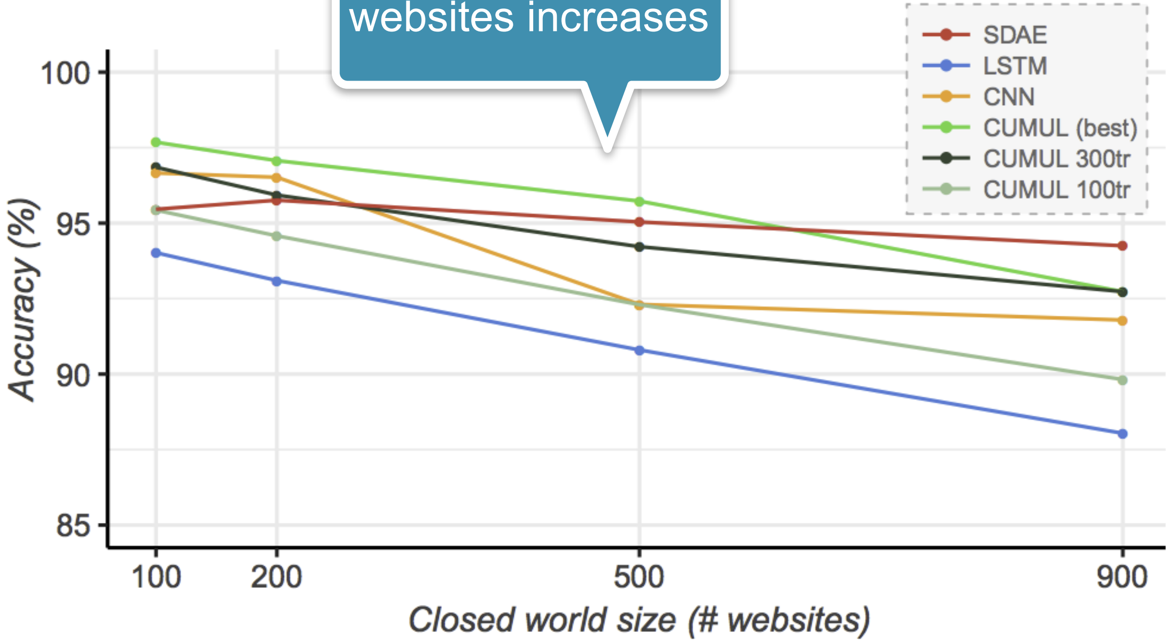
CW100: CUMUL still outperforms all attacks, followed by CNN



DistriNet

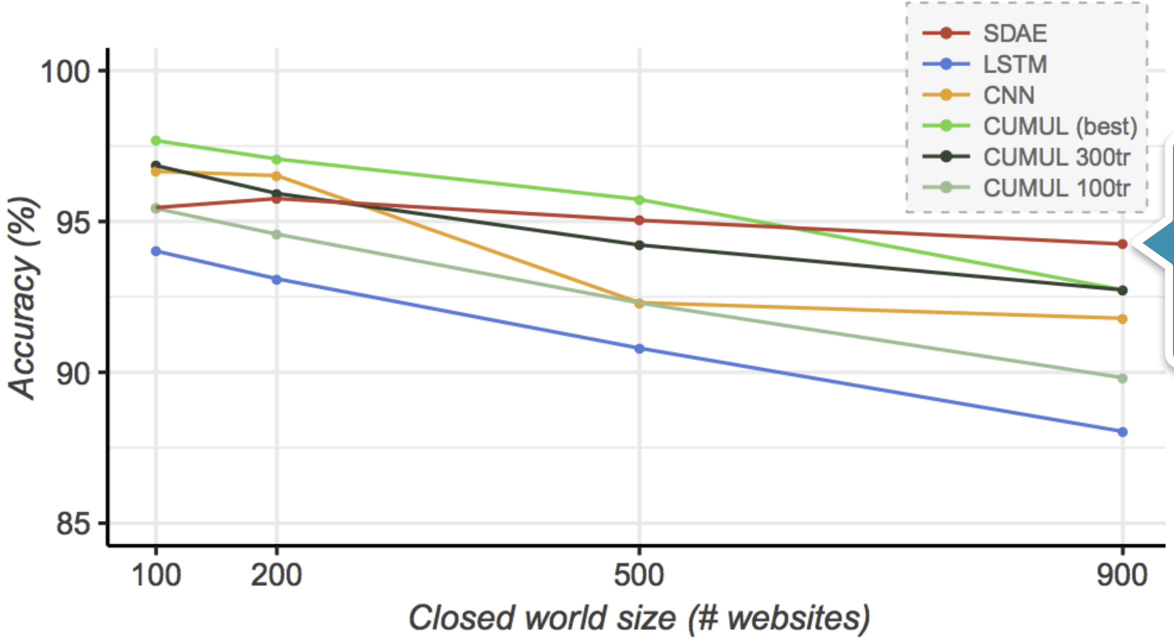
Closed World

Accuracy falls as the number of websites increases



DistriNet

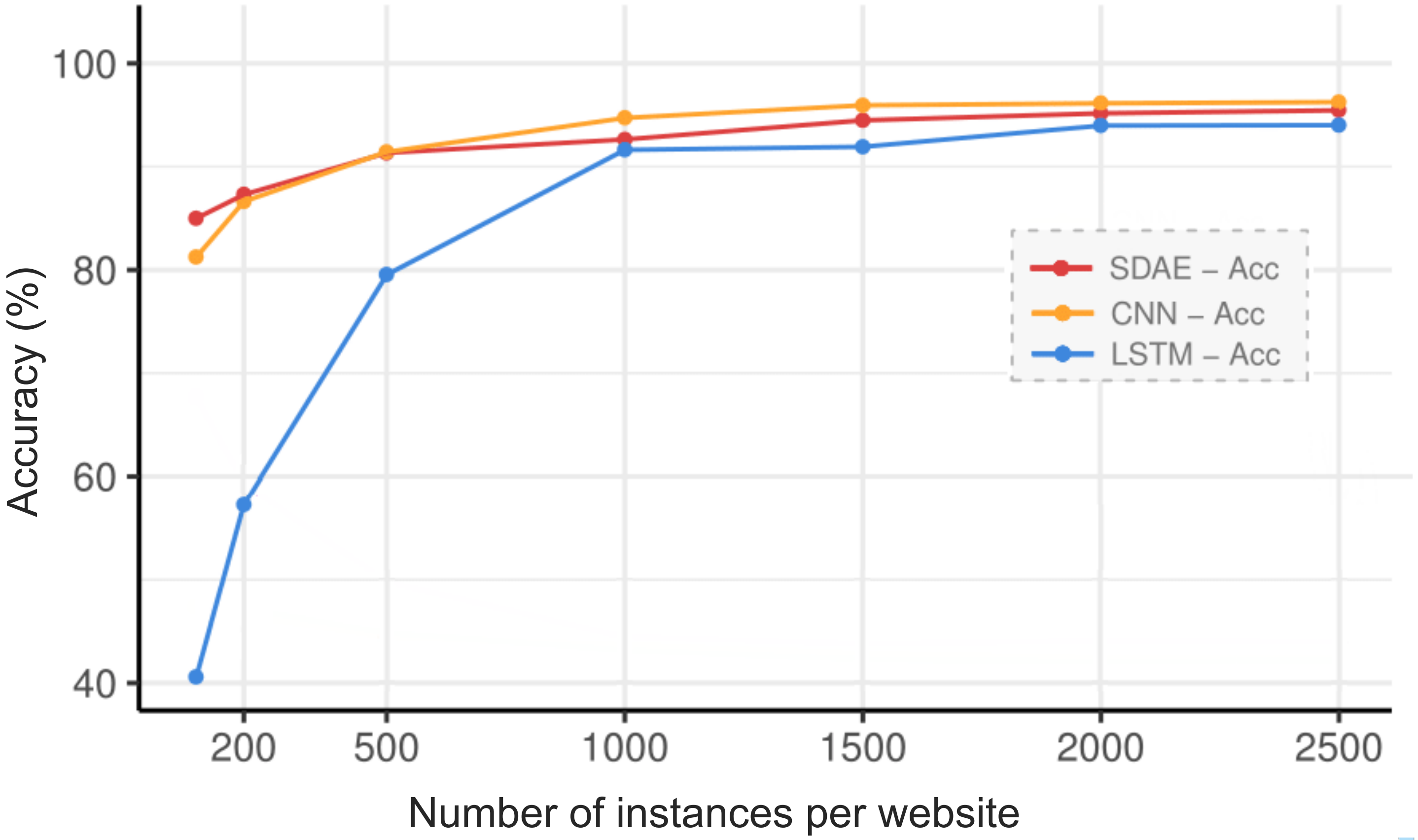
Closed World



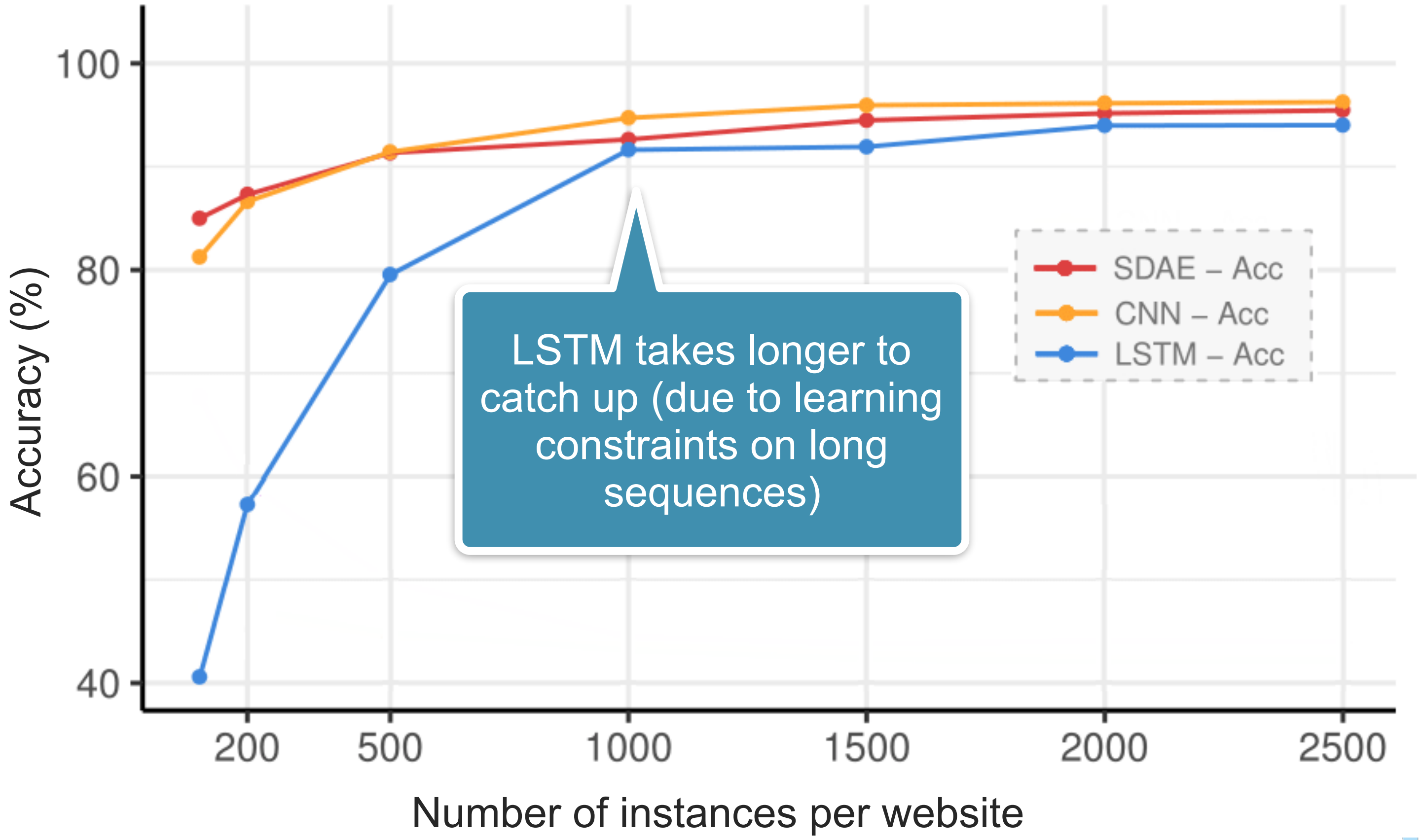
CW900: SDAE outperforms state-of-the-art

DistriNet

Number of Traces per Website

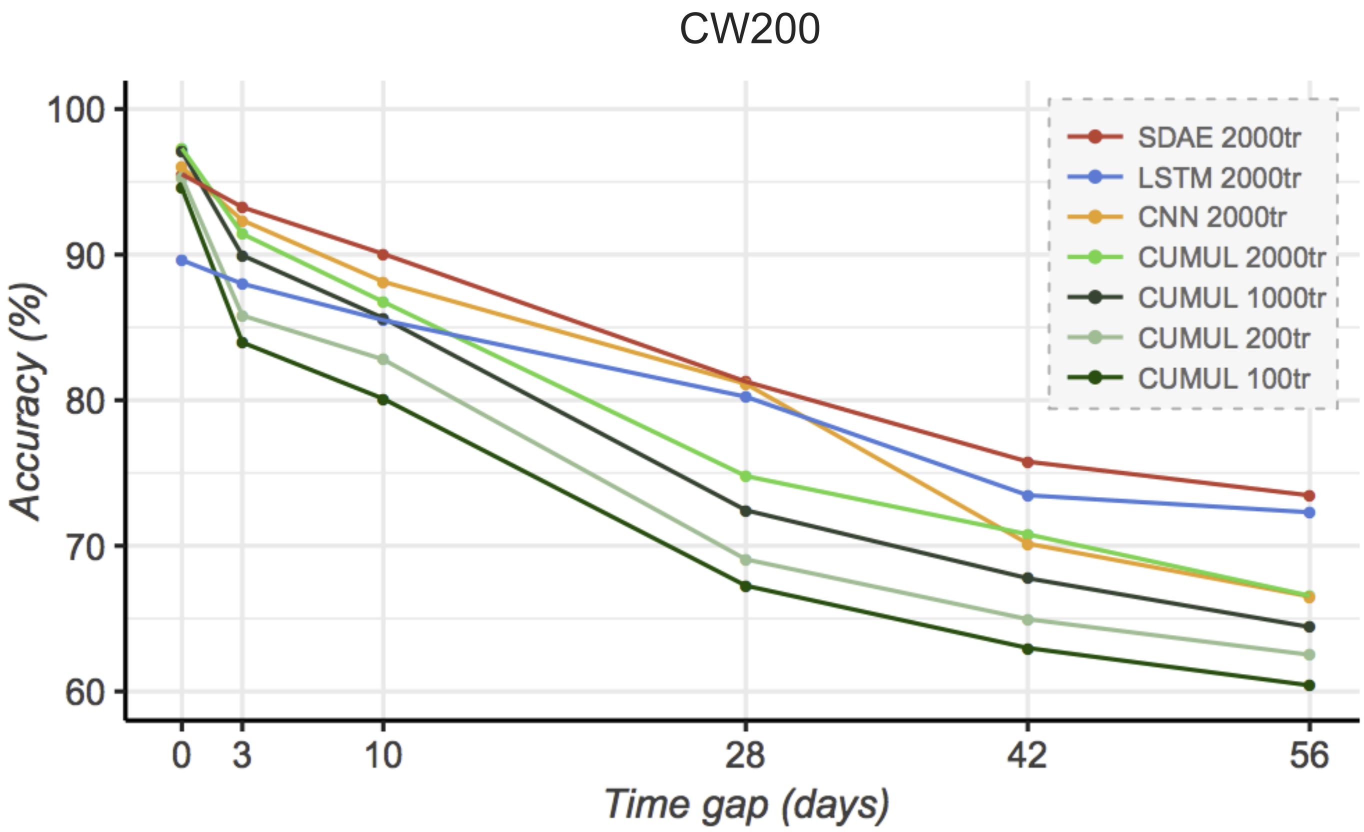


Number of Traces per Website

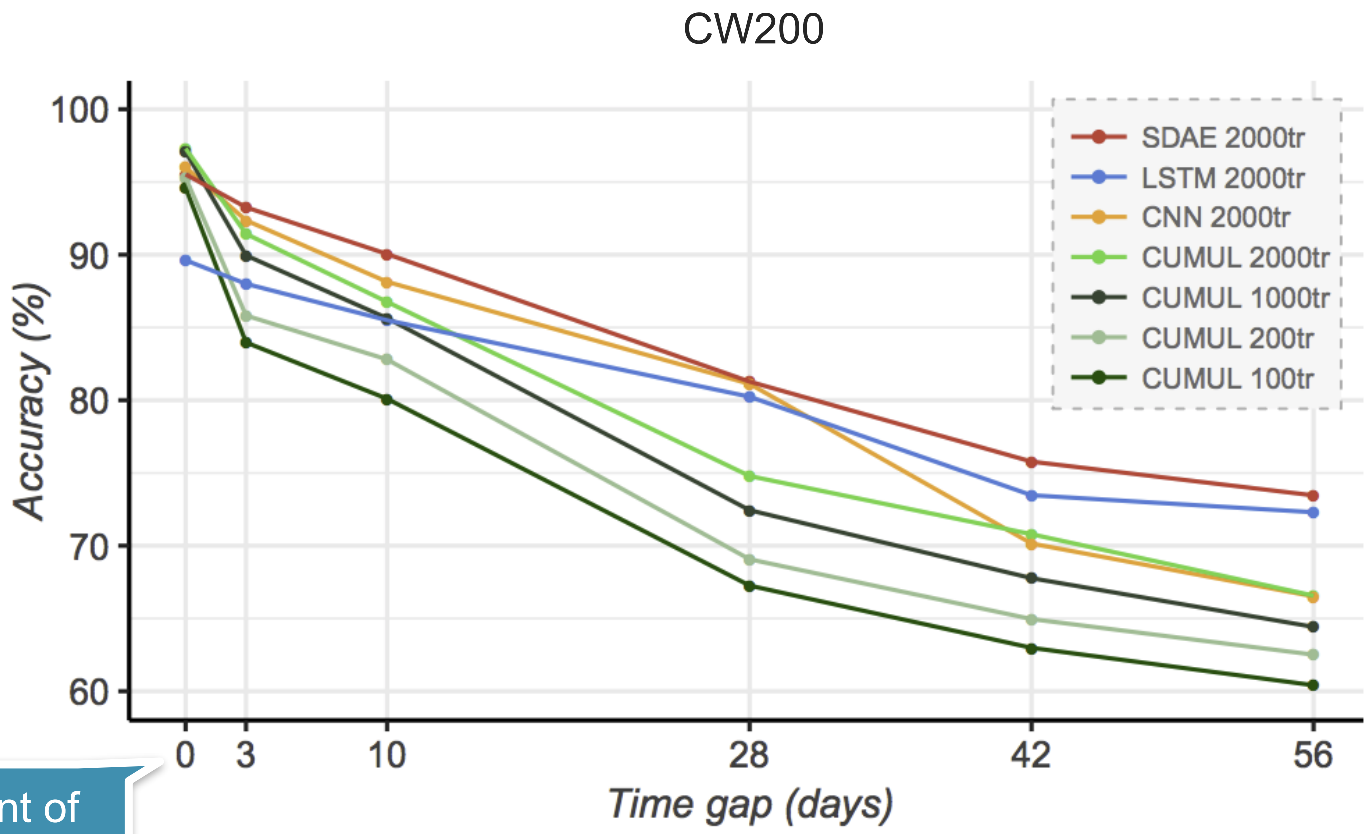


DistriNet

Concept Drift



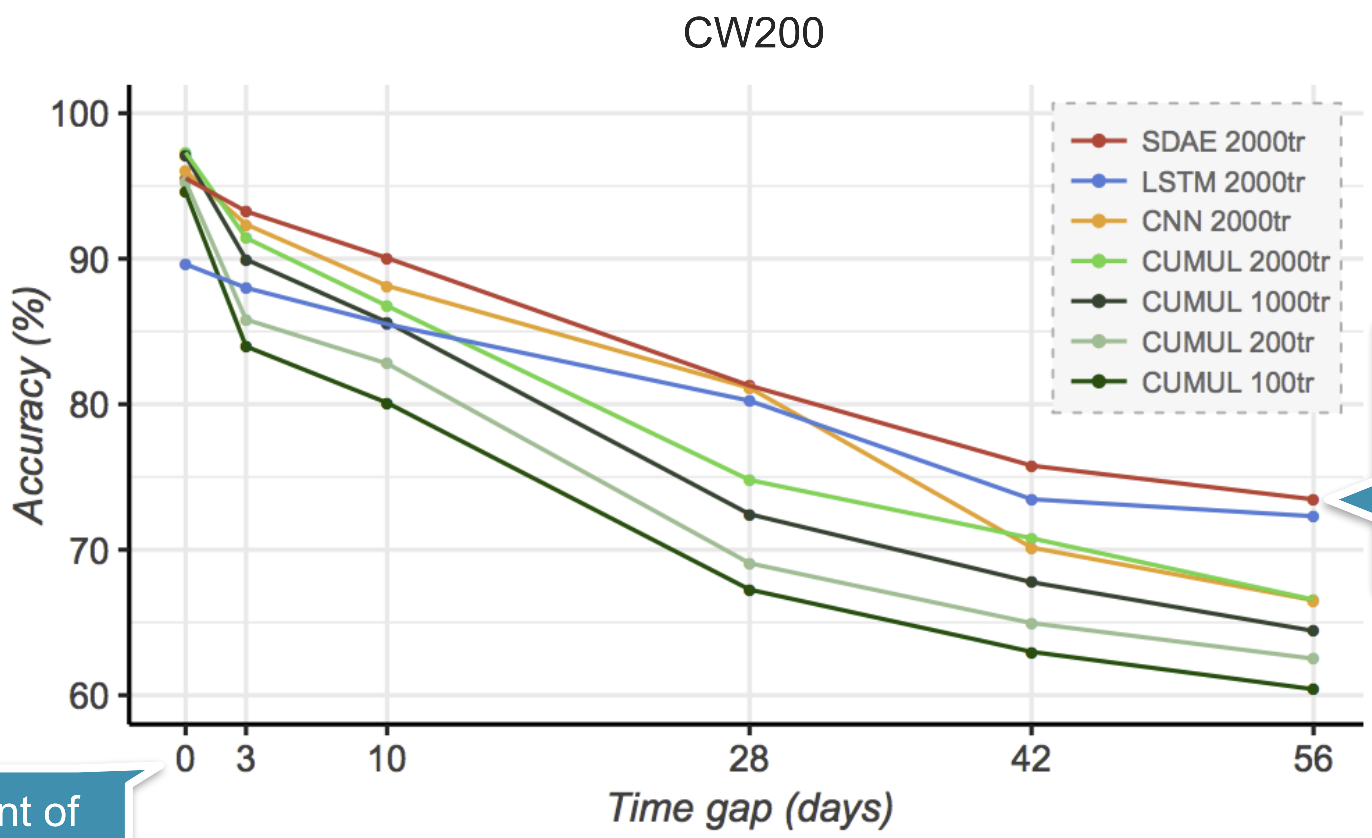
Concept Drift



Moment of training

DistriNet

Concept Drift



SDAE, LSTM and CNN generalize better than the state-of-the-art

Moment of training

DistriNet

Implications and Take-aways

Implications and Take-aways

- First thorough evaluation of DL for WF
  - Powerful and robust attack (accuracy: 96% for CW100, 94% for CW900)
  - Each DNN has its strengths and weaknesses
- Game-changer for the WF arms-race:
  - Automated feature learning (vs. the burden of manual feature engineering)
  - Harder to defend against (due to non-trivial interpretability of features)
- Data collection and model selection are crucial to the performance
  - Evaluated by collecting the largest dataset for WF

DistriNet

Thank you!

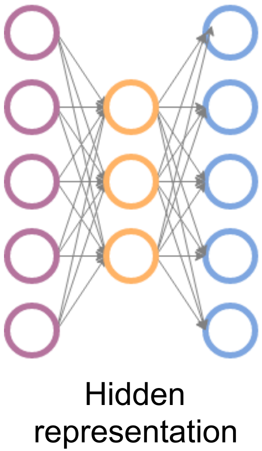
WEBSITE FINGERPRINTING THROUGH DEEP LEARNING

https://distrinet.cs.kuleuven.be/software/tor-wf-dl

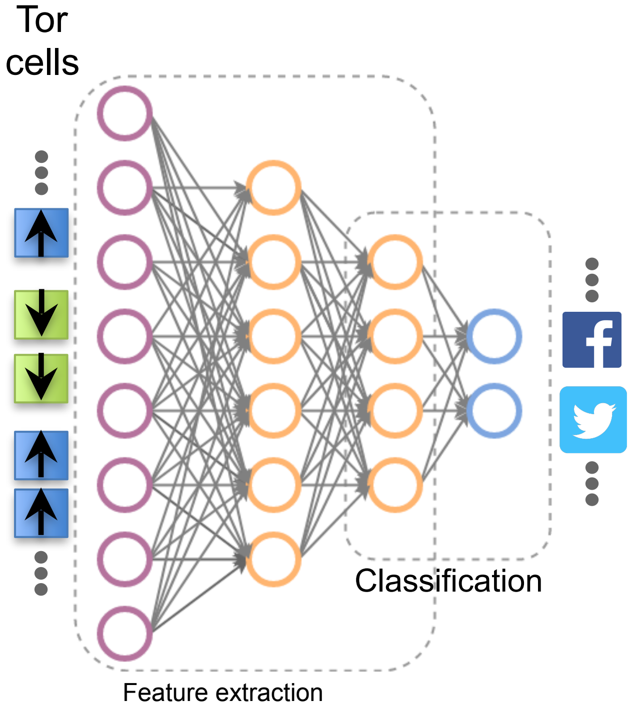
References

1. T. Wang and I. Goldberg, "Improved Website Fingerprinting on Tor," in ACM Workshop on Privacy in the Electronic Society (WPES). ACM, 2013, pp. 201–212.
2. T. Wang and I. Goldberg, "On realistically attacking tor with website fingerprinting," in Proceedings on Privacy Enhancing Technologies (PoPETs). De Gruyter Open, 2016, pp. 21–36.
3. T. Wang, X. Cai, R. Nithyanand, R. Johnson, and I. Goldberg, "Effective Attacks and Provable Defenses for Website Fingerprinting," in USENIX Security Symposium. USENIX Association, 2014, pp. 143–157.
4. A. Panchenko, F. Lanze, A. Zinnen, M. Henze, J. Pennekamp, K. Wehrle, and T. Engel, "Website fingerprinting at internet scale," in Network & Distributed System Security Symposium (NDSS). IEEE Computer Society, 2016, pp. 1–15.
5. J. Hayes and G. Danezis, "k-fingerprinting: a Robust Scalable Website Fingerprinting Technique," in USENIX Security Symposium. USENIX Association, 2016, pp. 1–17.
6. K. Abe and S. Goto, "Fingerprinting attack on tor anonymity using deep learning," Proceedings of the Asia-Pacific Advanced Network, vol. 42, pp. 15–20, 2016.
7. M. Juarez, S. Afroz, G. Acar, C. Diaz, and R. Greenstadt, "A critical evaluation of website fingerprinting attacks," in ACM Conference on Computer and Communications Security (CCS). ACM, 2014, pp. 263–274.

SDAE

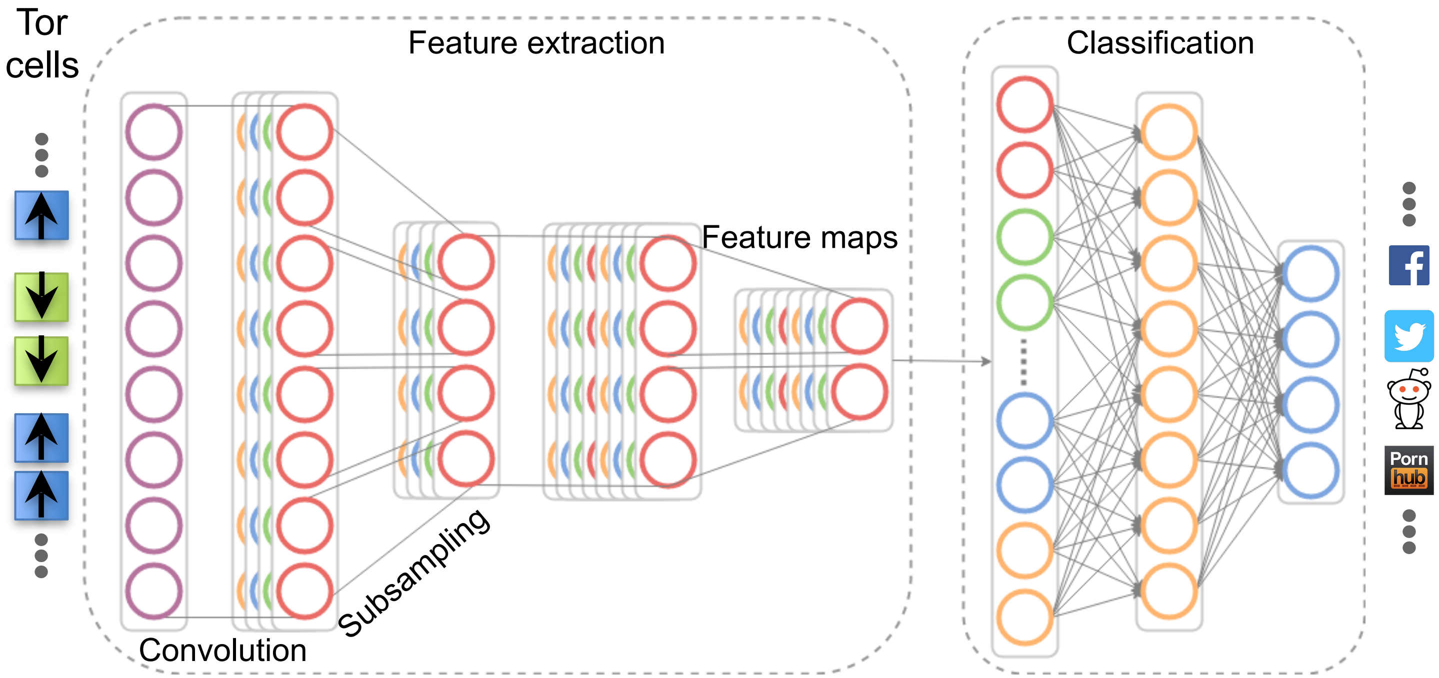


Autoencoder



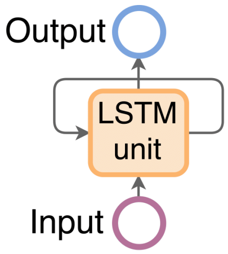
SDAE classifier

CNN

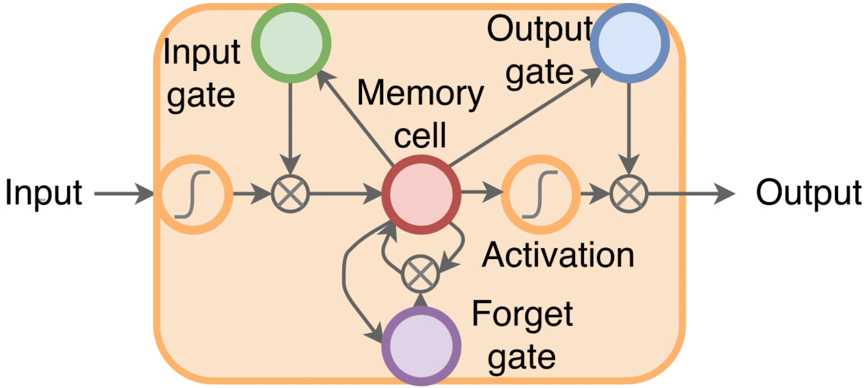


CNN classifier

LSTM

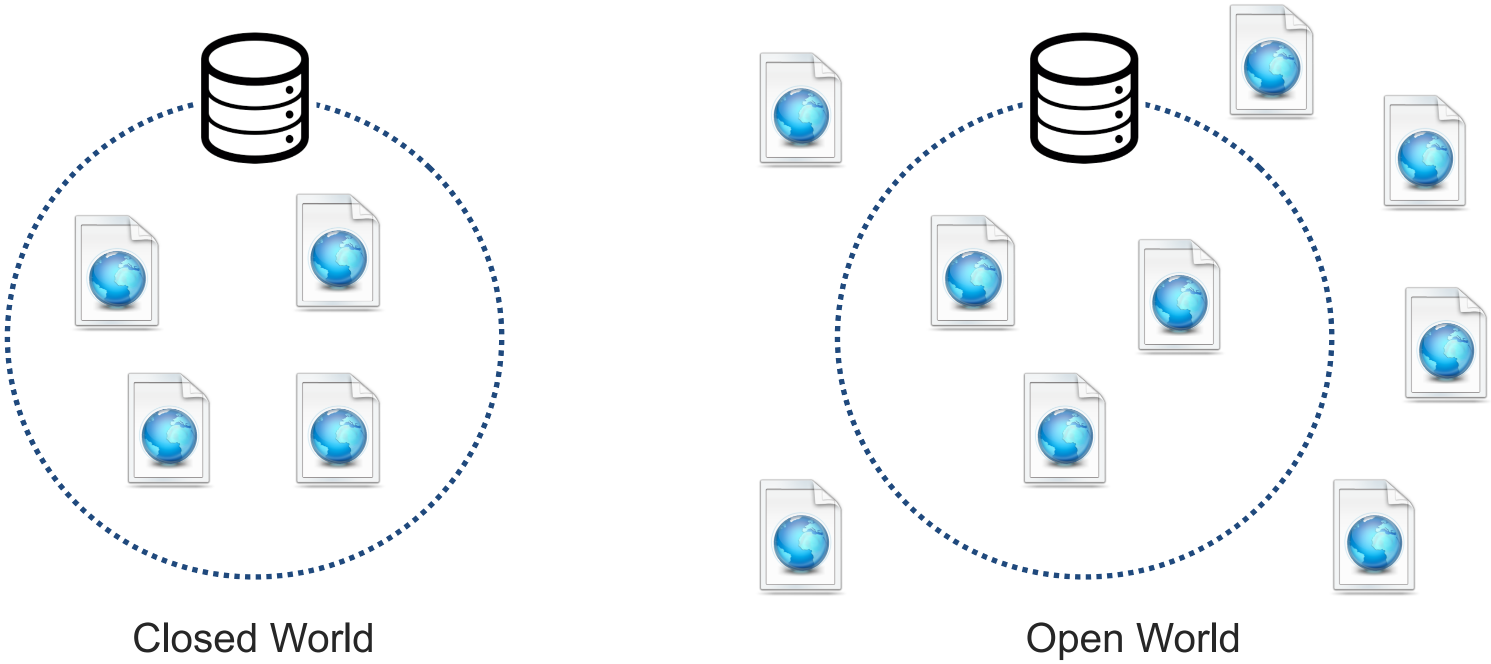


LSTM network



LSTM unit

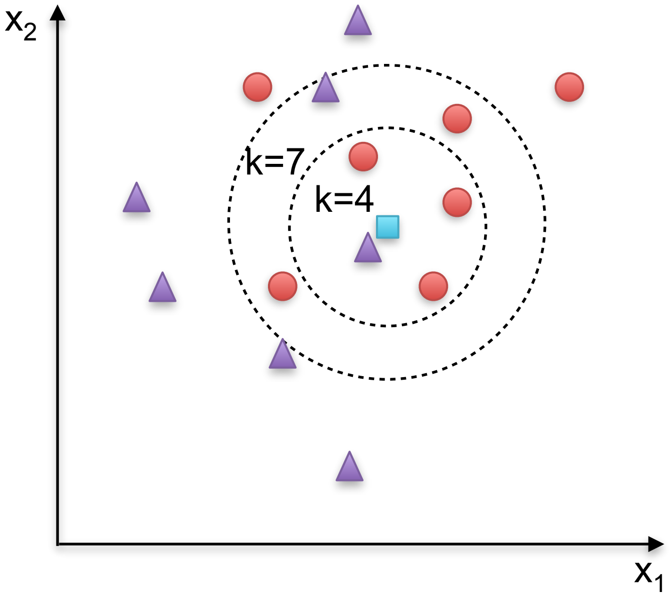
Closed World vs Open World



DistriNet

State-of-the-Art Attacks

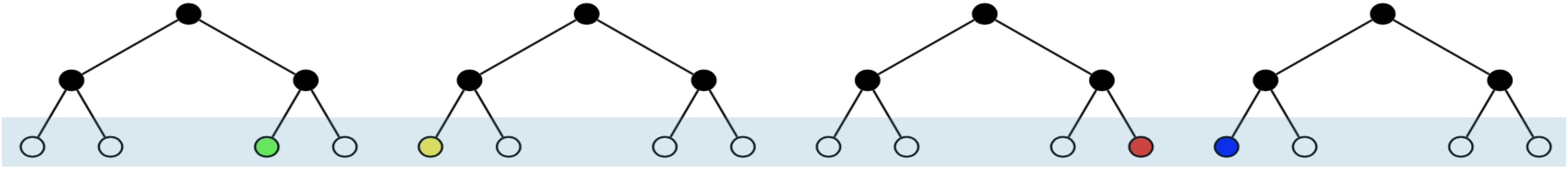
- kNN (Wang et al., 2014)
- Features
  - 3,000 (picked through heuristics)
  - total size, total time, number of packets, packet ordering, traffic bursts…
- Classifier
  - k-Nearest Neighbors (k-NN)
- Accuracy
  - 92% (100 websites)



DistriNet

State-of-the-Art Attacks

- k-Fingerprinting (Hayes et al, 2016)

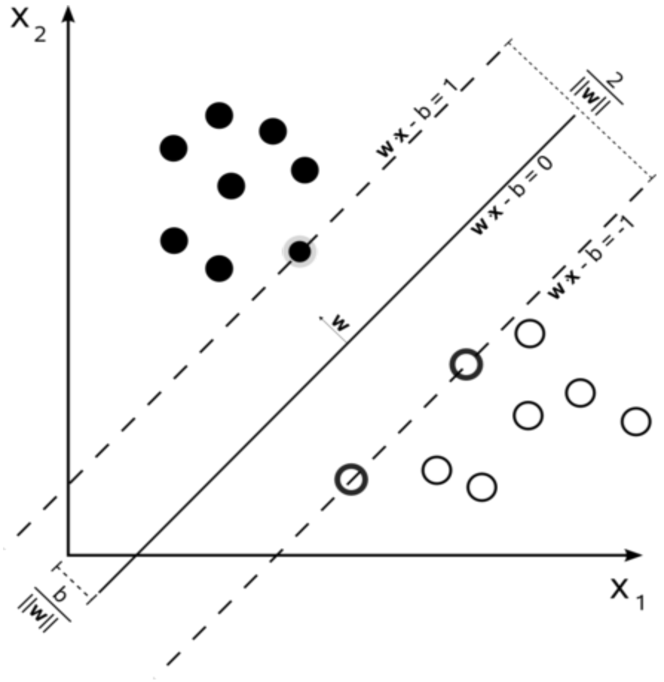


- Features
  - 150 (selected from Wang's through analysis of feature importance)
- Classifier
  - Random Forest + k-NN
- Accuracy
  - 93% (100 websites)

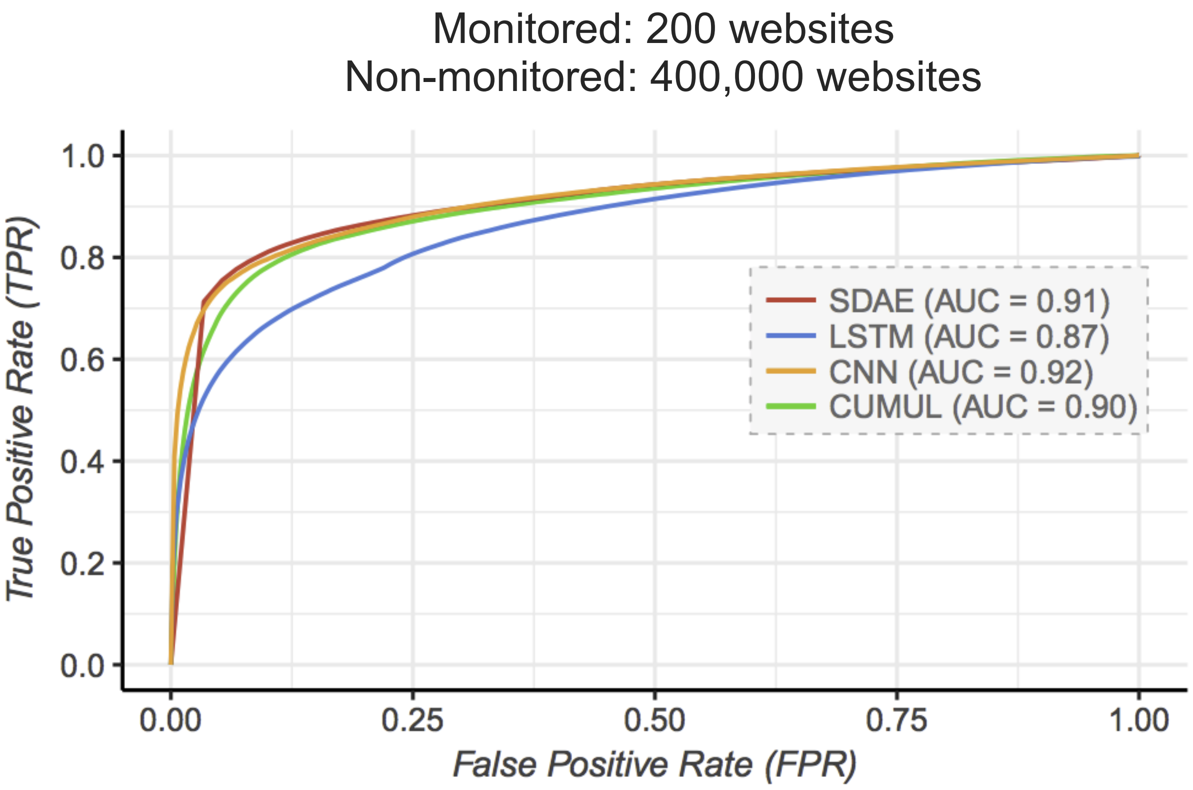
State-of-the-Art Attacks

CUMUL (Panchenko et al, 2016)

- Features
  - 100 (derived as interpolation points of the cumulative sum of packet lengths)
- Classifier
  - Support Vector Machine (SVM)
- Accuracy
  - From 97% (100 websites)



Open World: ROC Curve

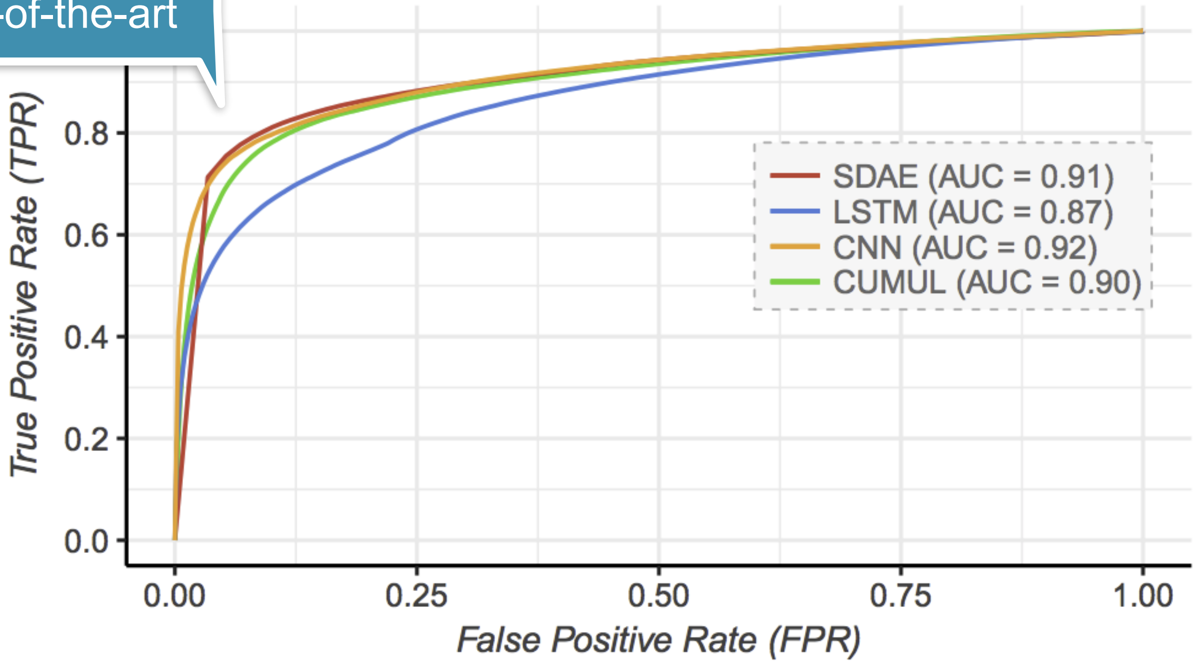


DistriNet

Open World: ROC Curve

CNN and SDAE outperform state-of-the-art

Monitored: 200 websites
Non-monitored: 400,000 websites



DistriNet