Data-driven models of reputation in cyber-security

Pierre Lison
Norwegian Computing Center (NR)

Tekna seminar,
Digital trussel-etterretning med AI

05/02/2019

Introduction

- Blacklists and whitelists (= **reputation lists**) often employed to filter network traffic
- Manually curated by security experts

Introduction

- **Shortcomings of blacklists and whitelists:**
  - Slow reaction time
  - Maintenance is difficult and time-consuming
  - Limited coverage
  - Static (can be circumvented through techniques such domain flux and fast flux networks)

Introduction

Can we use **machine learning** to automatically predict the reputation of end-point hosts?

Introduction

Can we use **machine learning** to automatically predict the reputation of end-point hosts?

1. Predictions in real-time, without human intervention
2. Less vulnerable to human errors and omissions
3. Full coverage of end-point hosts

Introduction

Can we use **machine learning** to automatically predict the reputation of end-point hosts?

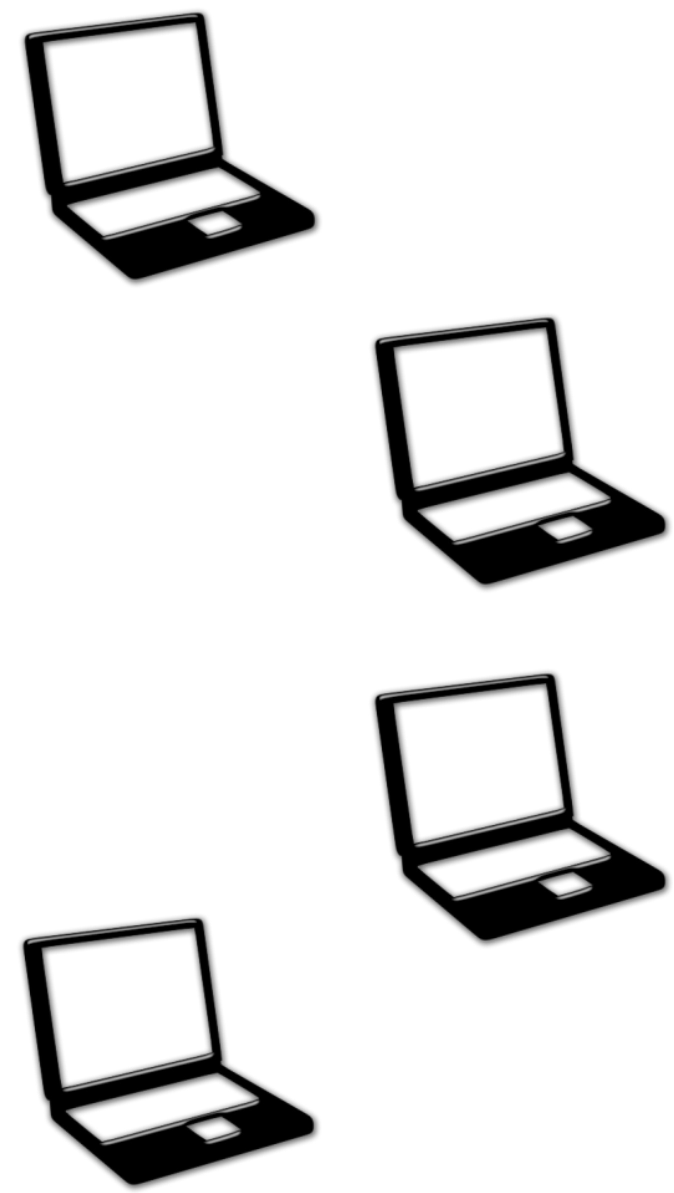
Detecting domain names generated by malware with RNNs

Predicting the reputation of domains and IP addresses from passive DNS data

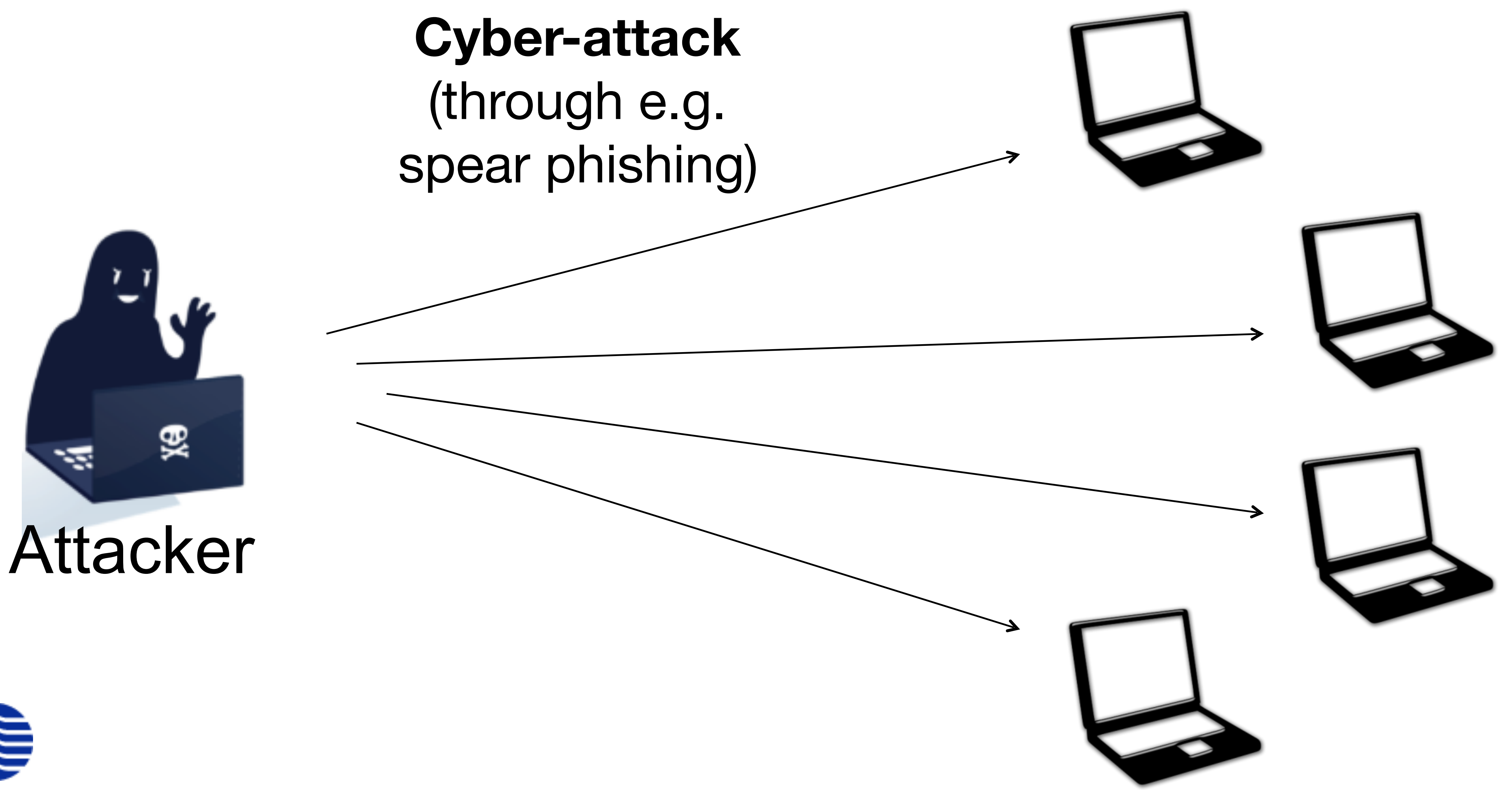
Part 1: Detecting domain names generated by malware

Domain-generating malware

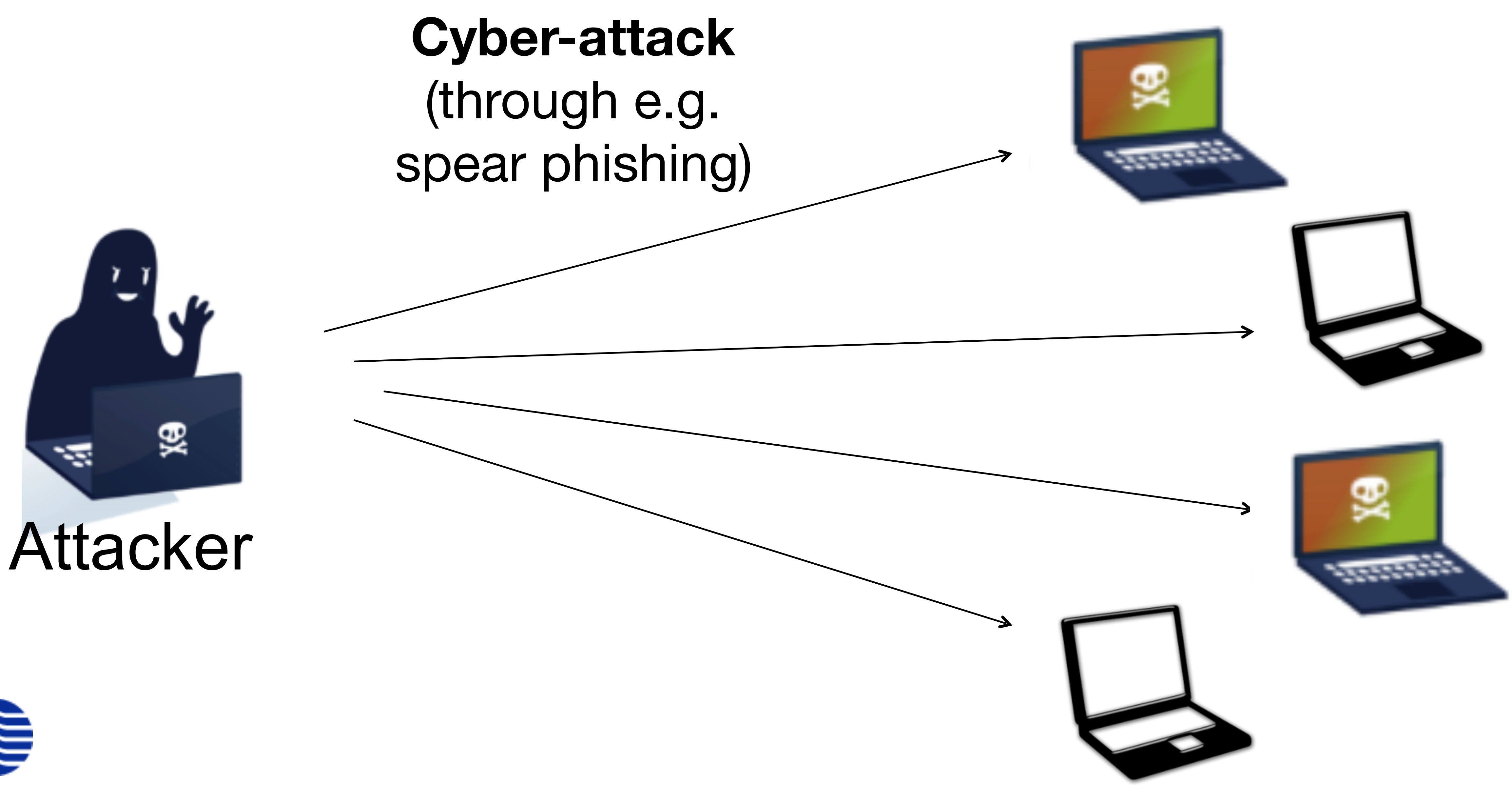
Attacker



Domain-generating malware

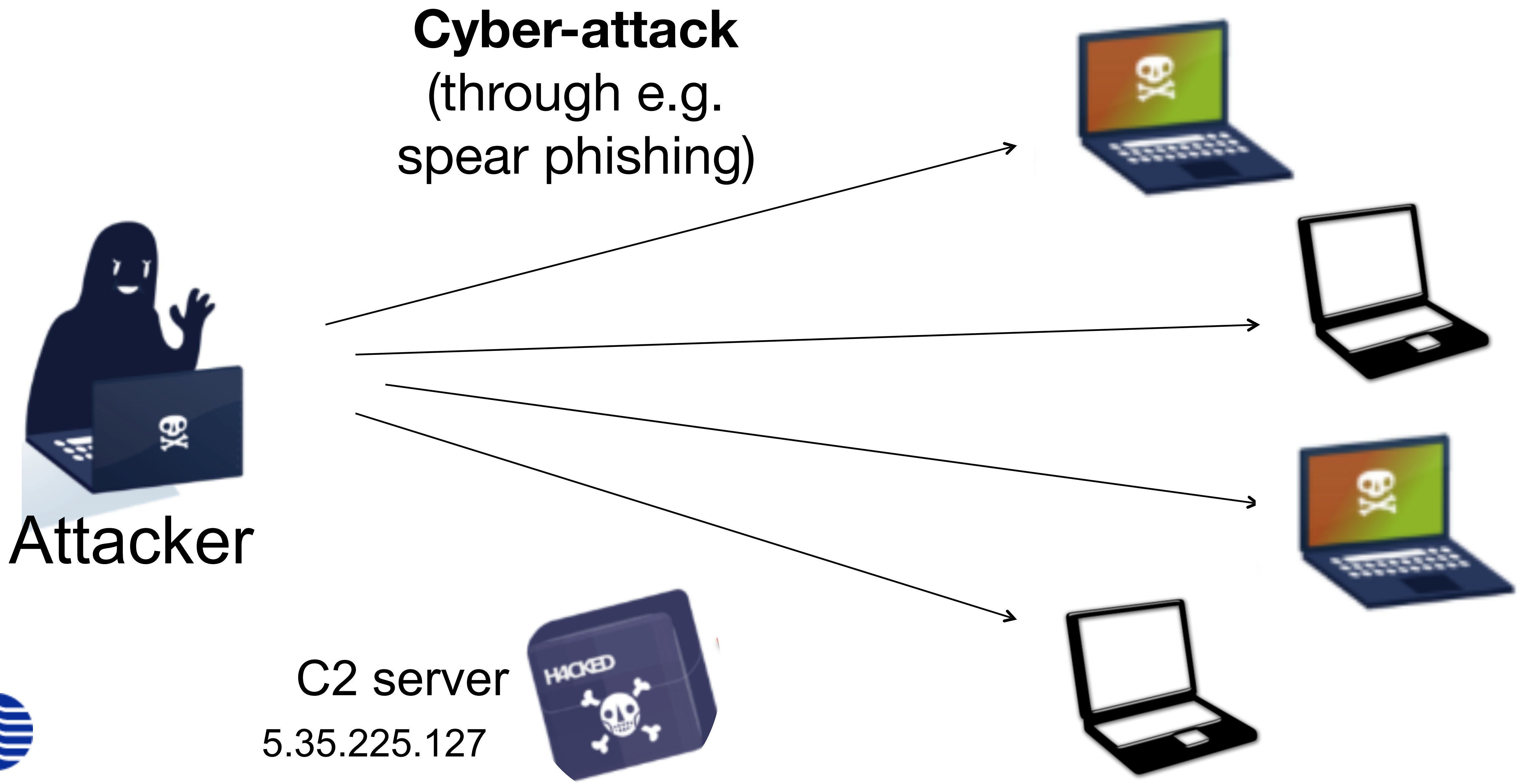


Domain-generating malware



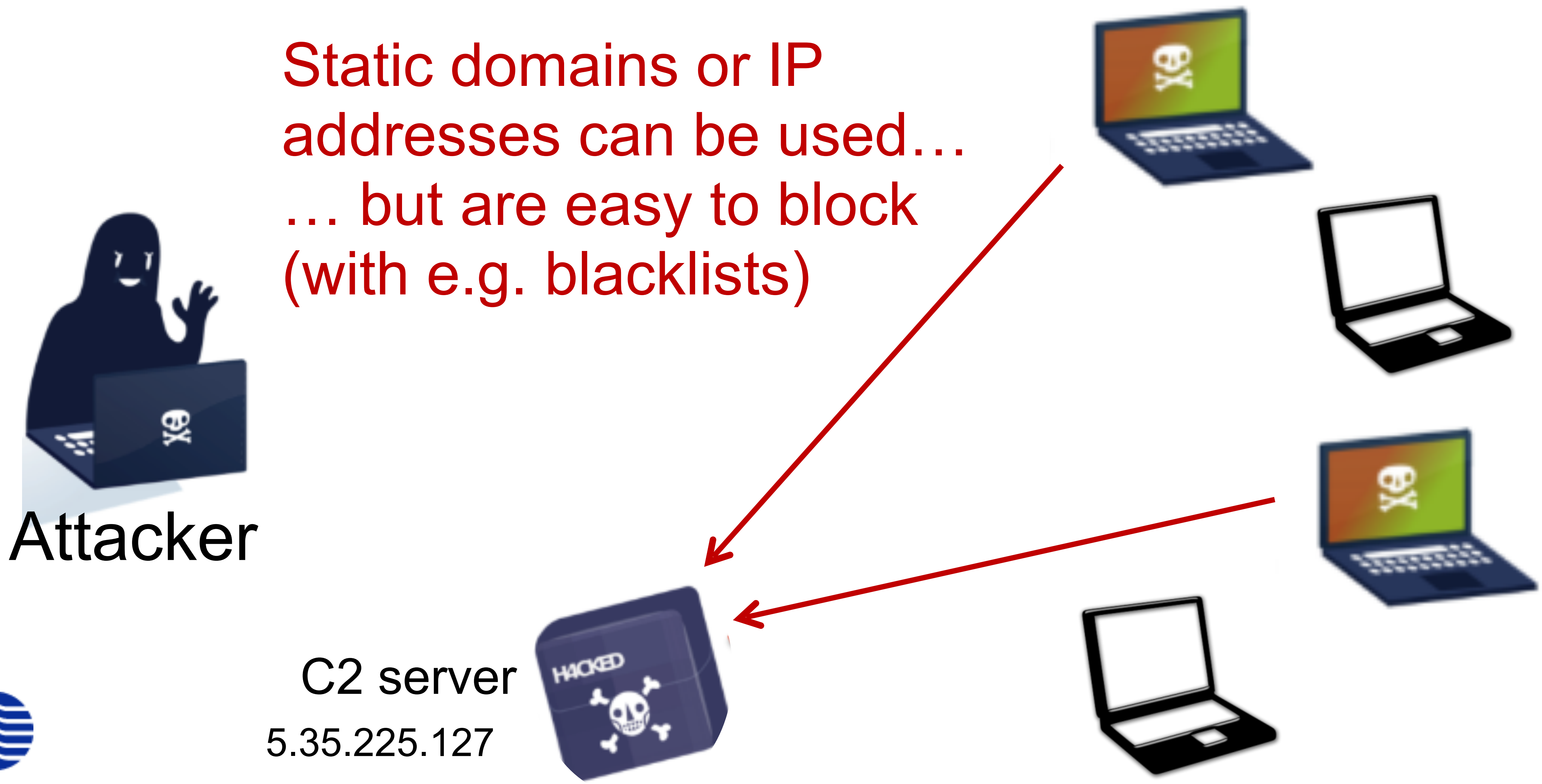
Domain-generating malware

- Most malware must connect compromised machines with a *command and control* (C2) server for their operations



Domain-generating malware

- Most malware must connect compromised machines with a *command and control* (C2) server for their operations



Domain-generating malware

- With *domain-generation algorithms* (DGA), compromised machines will attempt to connect to a large number of pseudo-random domain names…

Attacker

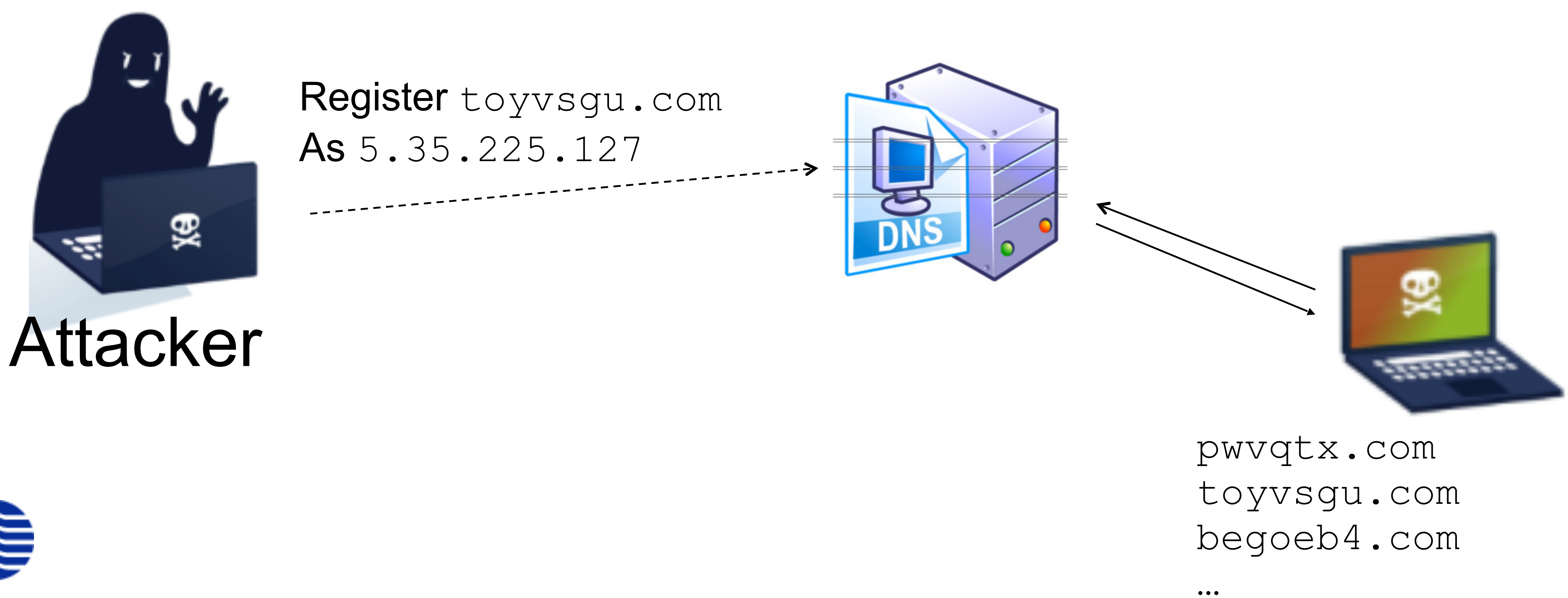
Domain-generating malware

- With *domain-generation algorithms* (DGA), compromised machines will attempt to connect to a large number of pseudo-random domain names…

pwvqtx.com
toyvsgu.com
begoeb4.com
…

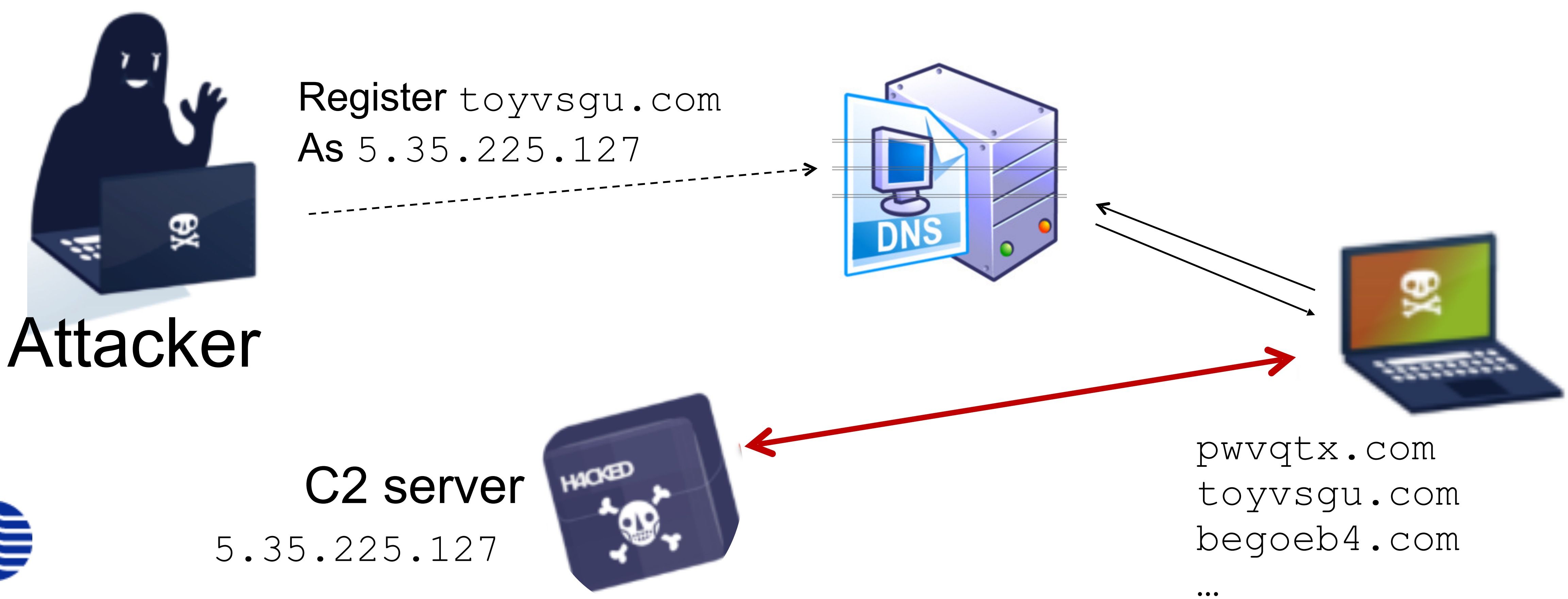
Domain-generating malware

- With *domain-generation algorithms* (DGA), compromised machines will attempt to connect to a large number of pseudo-random domain names…
- The attacker can then simply register a few of these artificial domains to establish a *rendez-vous point*



Domain-generating malware

- With *domain-generation algorithms* (DGA), compromised machines will attempt to connect to a large number of pseudo-random domain names…
- The attacker can then simply register a few of these artificial domains to establish a *rendez-vous point*



Domain-generating algorithms (DGAs)

- Very popular rendez-vous mechanism
  - First observed in the Kraken botnet (2008)

Domain-generating algorithms (DGAs)

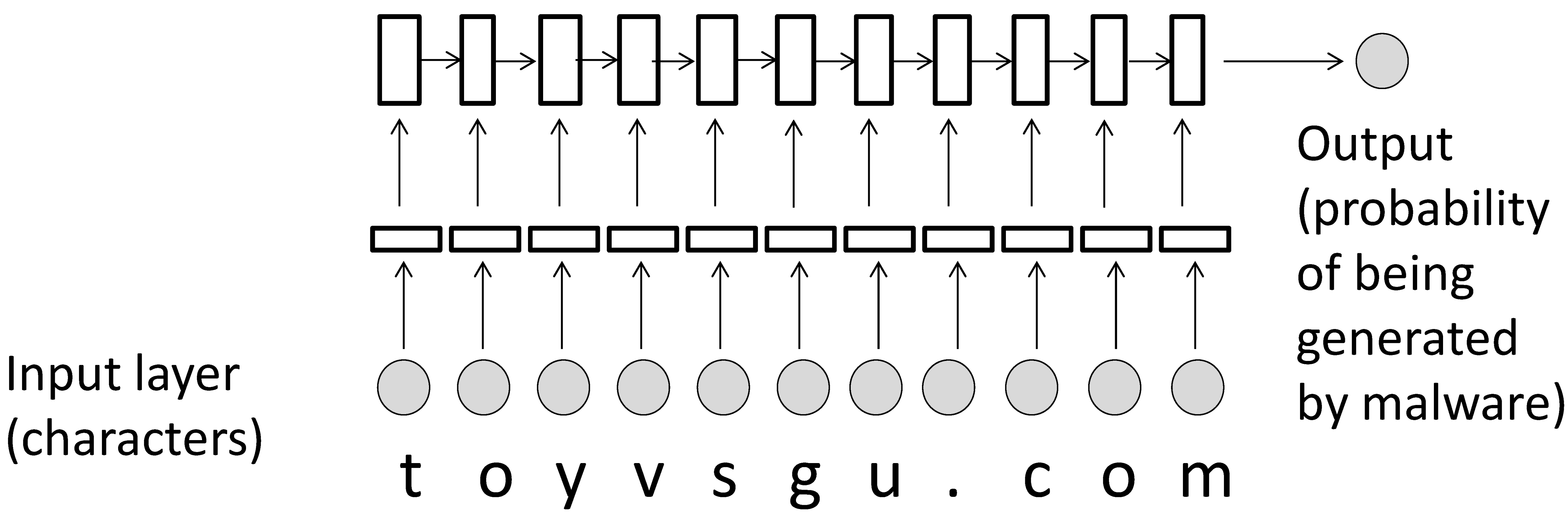
- Very popular rendez-vous mechanism
  - First observed in the Kraken botnet (2008)
- DGAs generate a large number of seemingly random domain names based on a *shared secret* (**seed**)
  - Various generation procedures (hash-based techniques, permutations, wordlists, etc.)
  - Static or time-dependent? Deterministic or stochastic?

Domain-generating algorithms (DGAs)

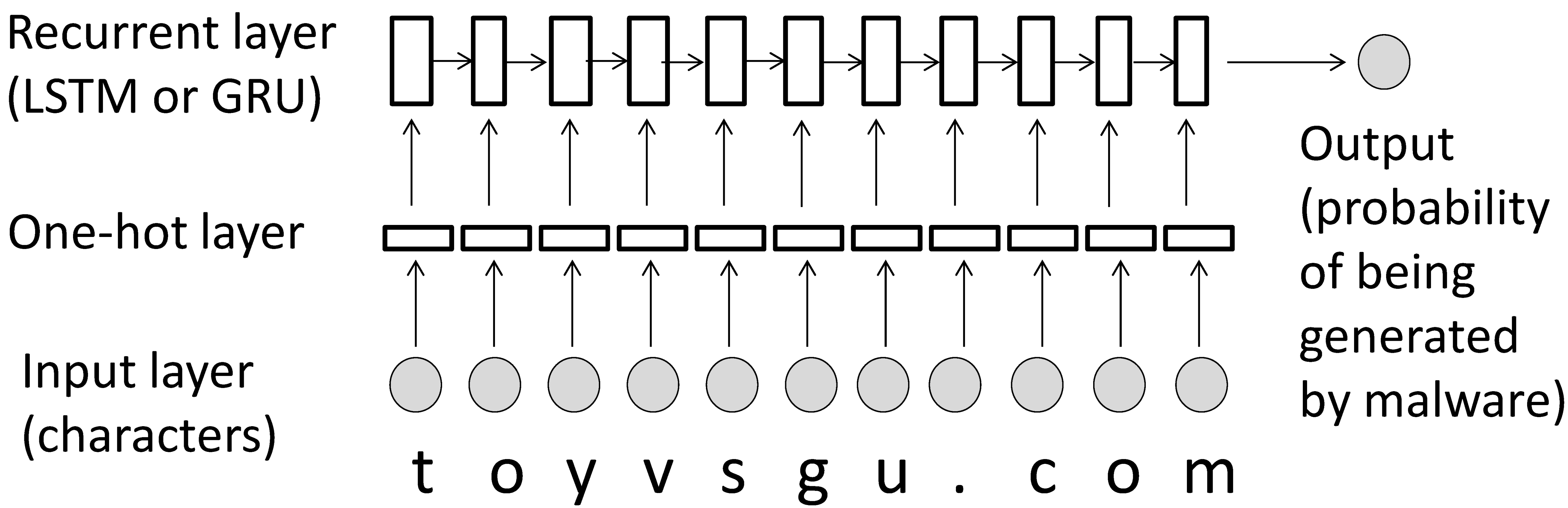
- Very popular rendez-vous mechanism
  - First observed in the Kraken botnet (2008)
- DGAs generate a large number of seemingly random domain names based on a *shared secret* (**seed**)
  - Various generation procedures (hash-based techniques, permutations, wordlists, etc.)
  - Static or time-dependent? Deterministic or stochastic?
- Highly **asymmetric** situation between malicious actors and security professionals

Detection of DGAs

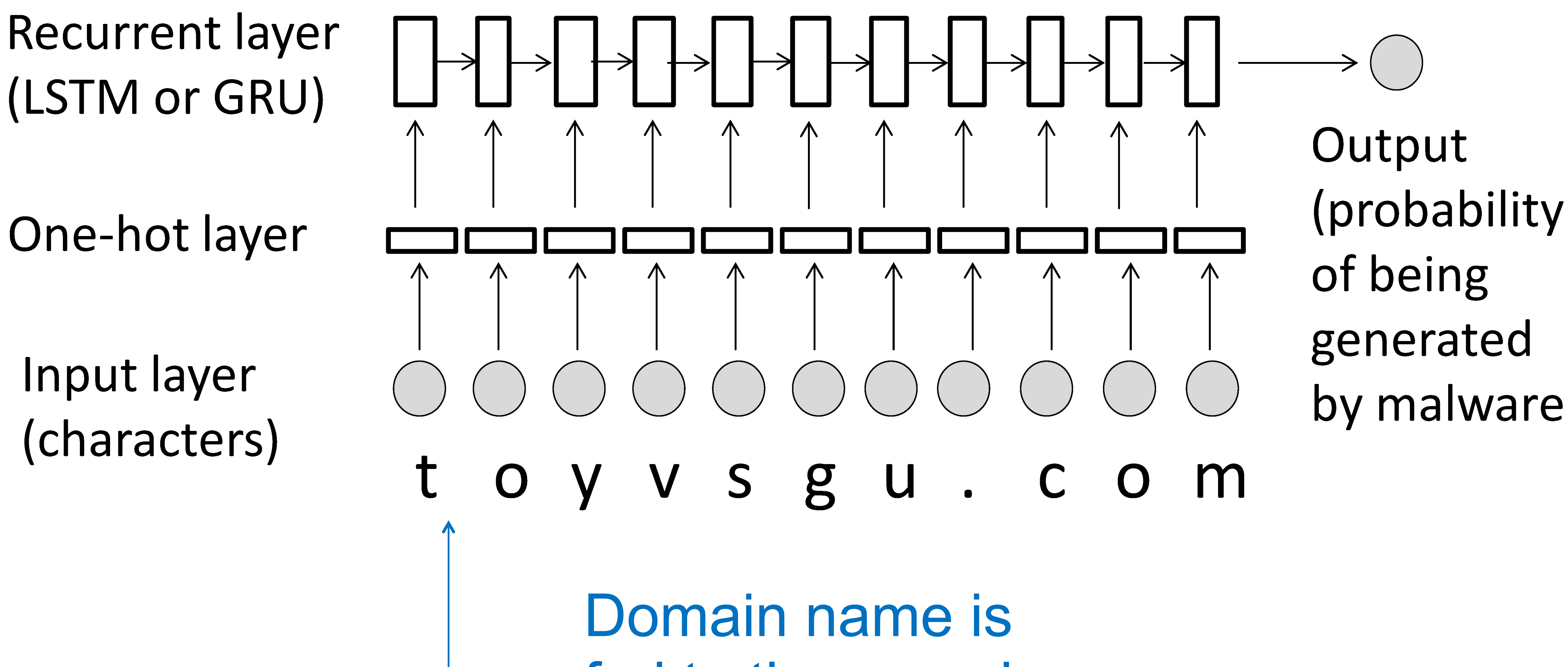
- **Recurrent neural network** trained on a large dataset of benign & malicious domains
  - Ability to learn complex sequential patterns
- Purely data-driven – easy to apply and update



Architecture

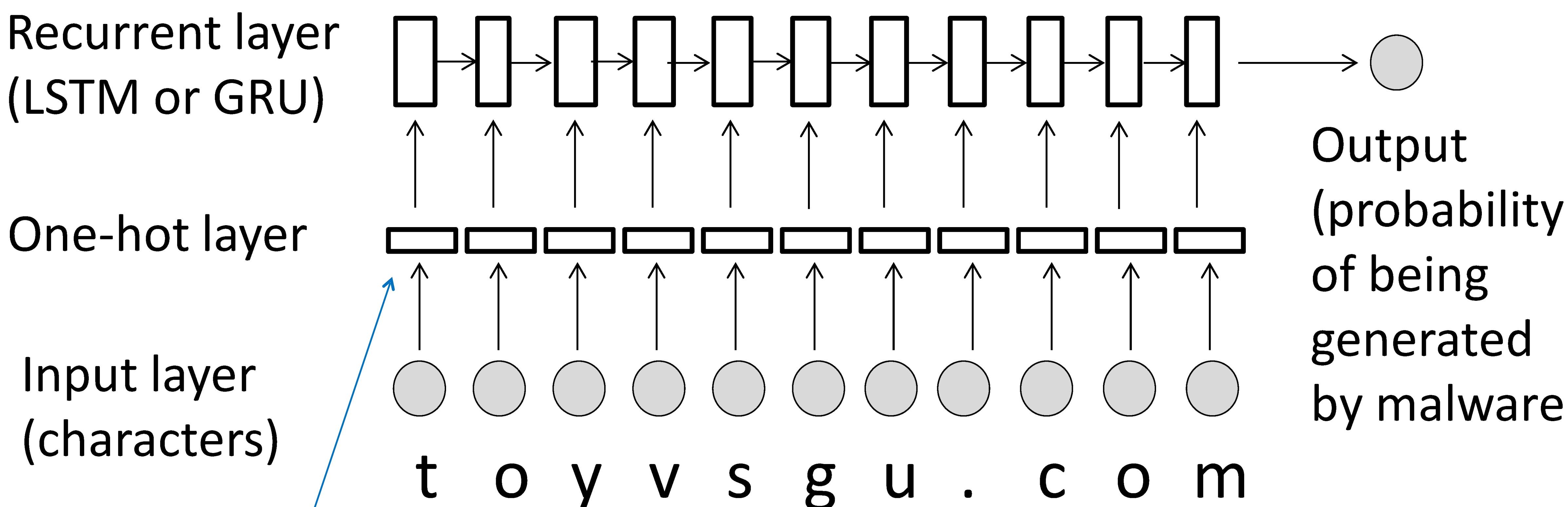


Architecture



Domain name is fed to the neural network character by character

Architecture

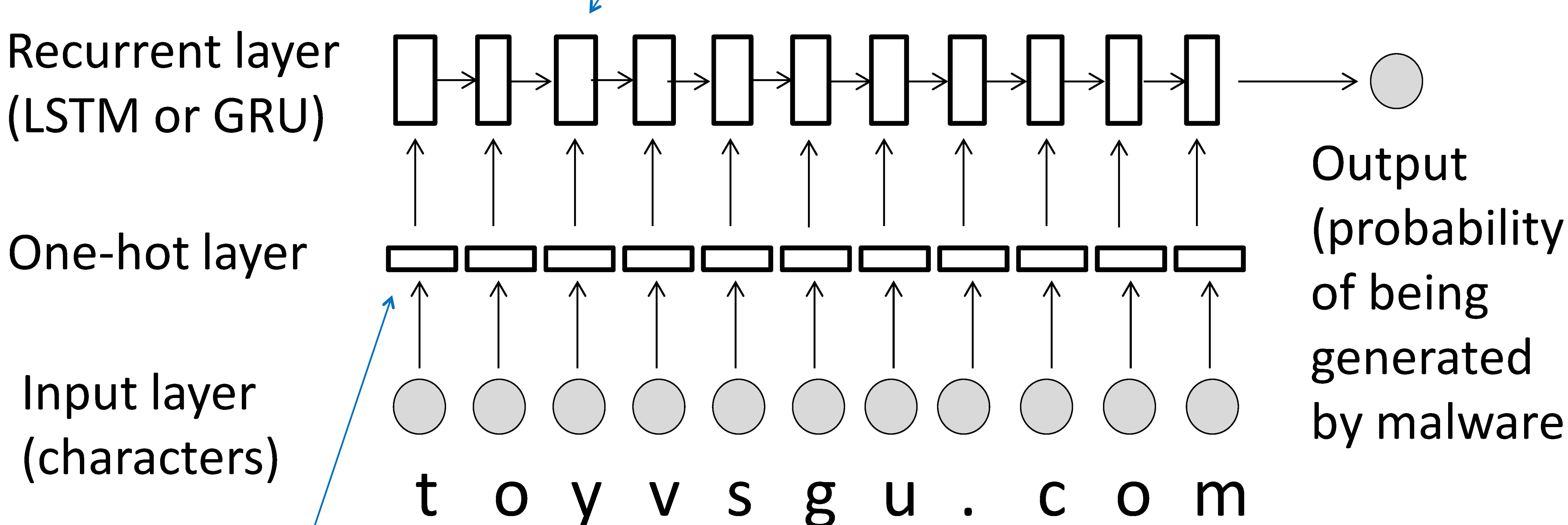


First layer encode each character as a "one-hot" vector

Domain name is fed to the neural network character by character

Architecture

Recurrent layer builds up a representation of the character sequence as a dense vector

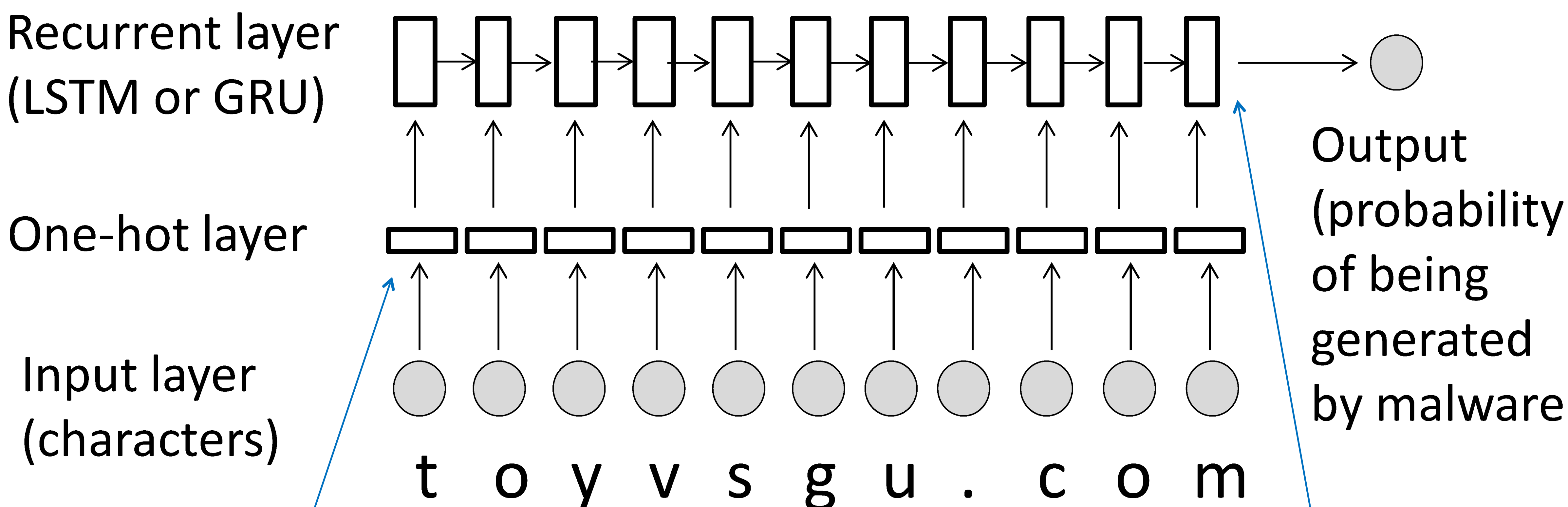


First layer encode each character as a "one-hot" vector

Domain name is fed to the neural network character by character

Architecture

Recurrent layer builds up a representation of the character sequence as a dense vector



First layer encode each character as a "one-hot" vector

Domain name is fed to the neural network character by character

Final vector is used to predict whether the domain is DGA

Data

- **Negative examples** (benign domains):
  - Snapshots from the Alexa top 1 million domains
  - Total: over 4 million domains

- **Positive examples** (malware DGAs)
  - DGArchive (63 types of malware)
  - Feeds from Bambenek Consulting
  - Domain generators for 11 DGAs
  - Total: 2.9 million domains

Results

[Lison, P., & Mavroeidis, V. (2017). Automatic Detection of Malware-Generated Domains with Recurrent Neural Models. In *Proceedings of NISK 2017*.]

- Detection

	Accuracy	Precision	Recall
Bigram	0.915	0.927	0.882
Neural model	**0.973**	**0.972**	**0.970**

- Classification

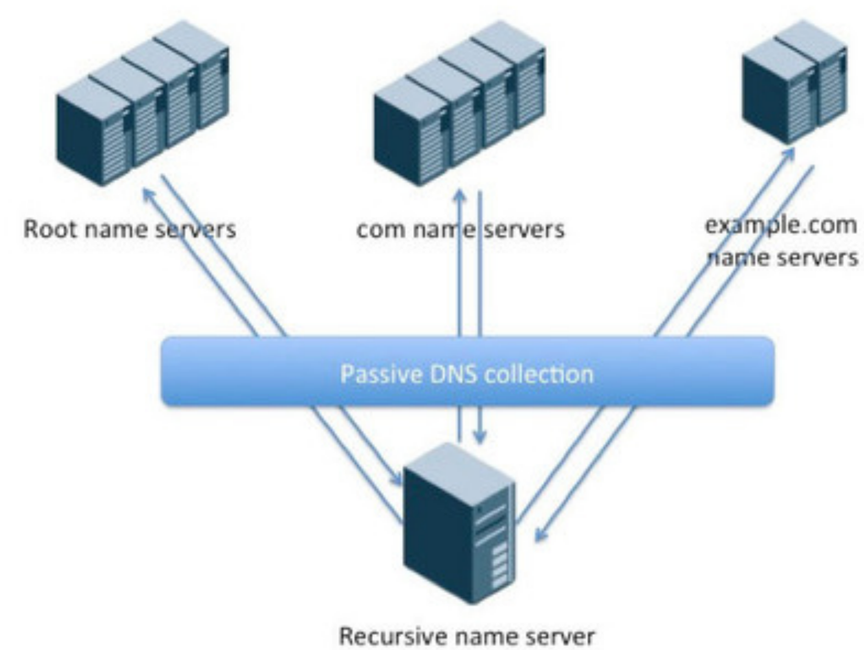
	Accuracy	Precision	Recall
Bigram	0.800	0.787	0.800
Neural model	**0.892**	**0.891**	**0.892**

Part 2: Predicting the reputation from passive DNS data

Passive DNS

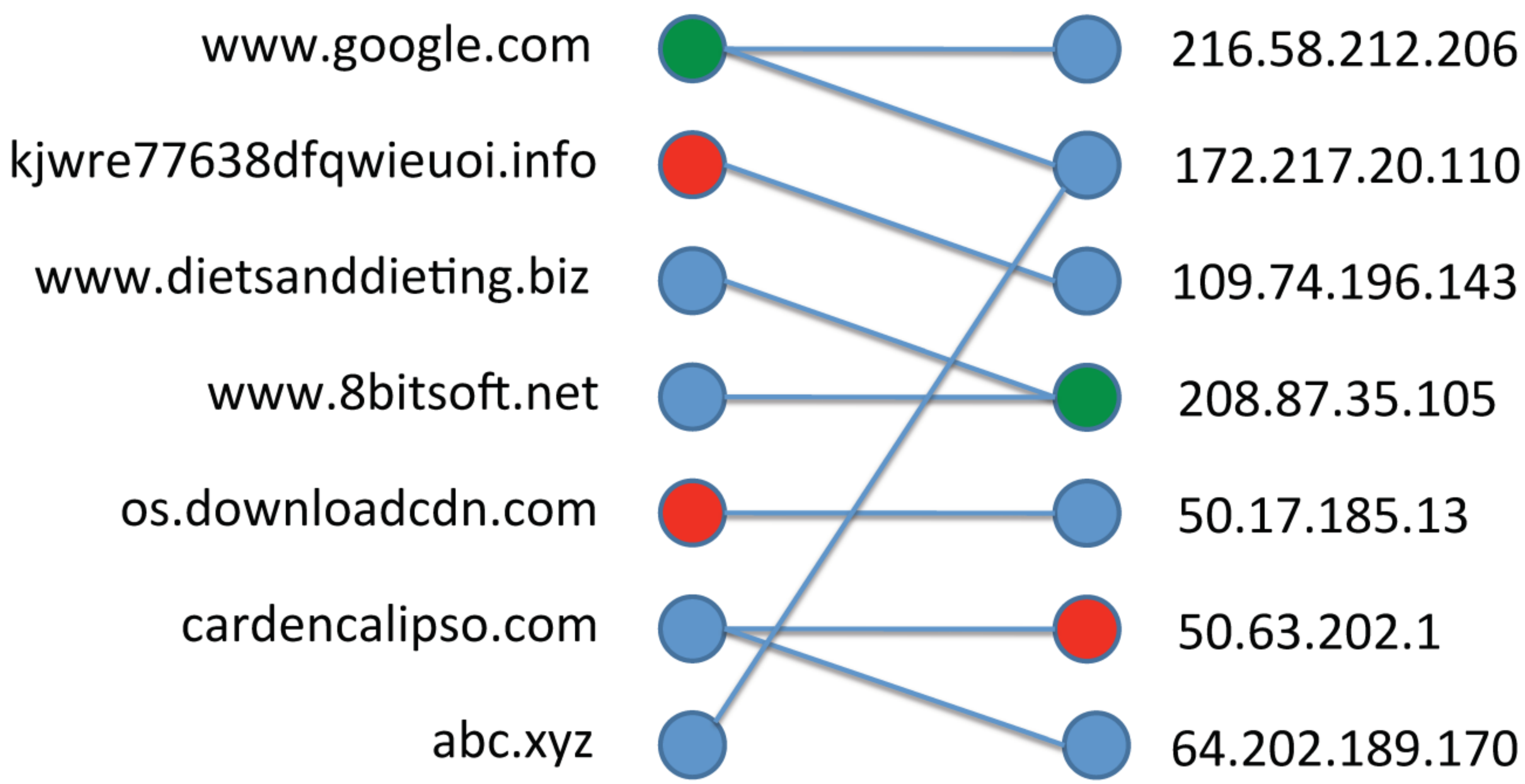
- **Passive DNS data** very useful for threat intelligence:
  - Inter-server DNS messages captured by sensors
  - Less privacy concerns (not tied to personal information)

- We used a dataset of *720 million aggregated DNS queries*
  - Covers a period of 4 years
  - Courtesy of Mnemonic AS [www.mnemonic.no]



Data

Labelled dataset of **720 million** records (**102 M** records labelled as benign, **8.2 M** records as malicious and **614 K** records as sinkhole)



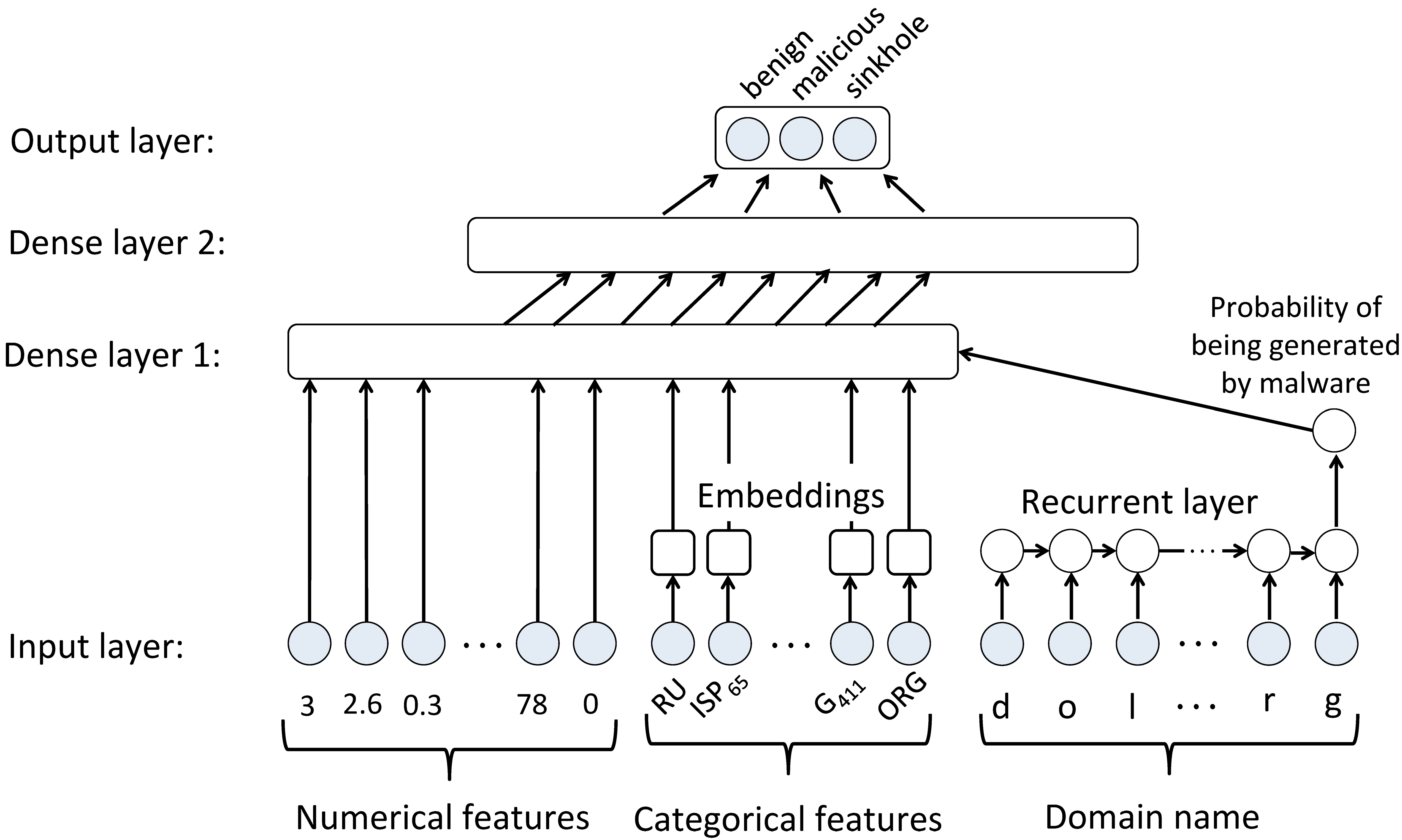
We enriched the passive DNS data with:

- Reputation labels from existing blacklists and whitelists
- IP location(geoname identifiers) and ISP data

Features

- Numerical features derived from the records:
  - Lifespan, number of queries (for record, domain or IP), number of distinct countries or ISP, TTL values, etc.
- Categorical features:
  - ISP, geolocation, top-level domain, etc.
- Ranking features from Alexa
- Features extracted from **graph inference**
  - Number of records at distance *n* and of reputation *X*
- Sequence of characters from the domain

Neural model

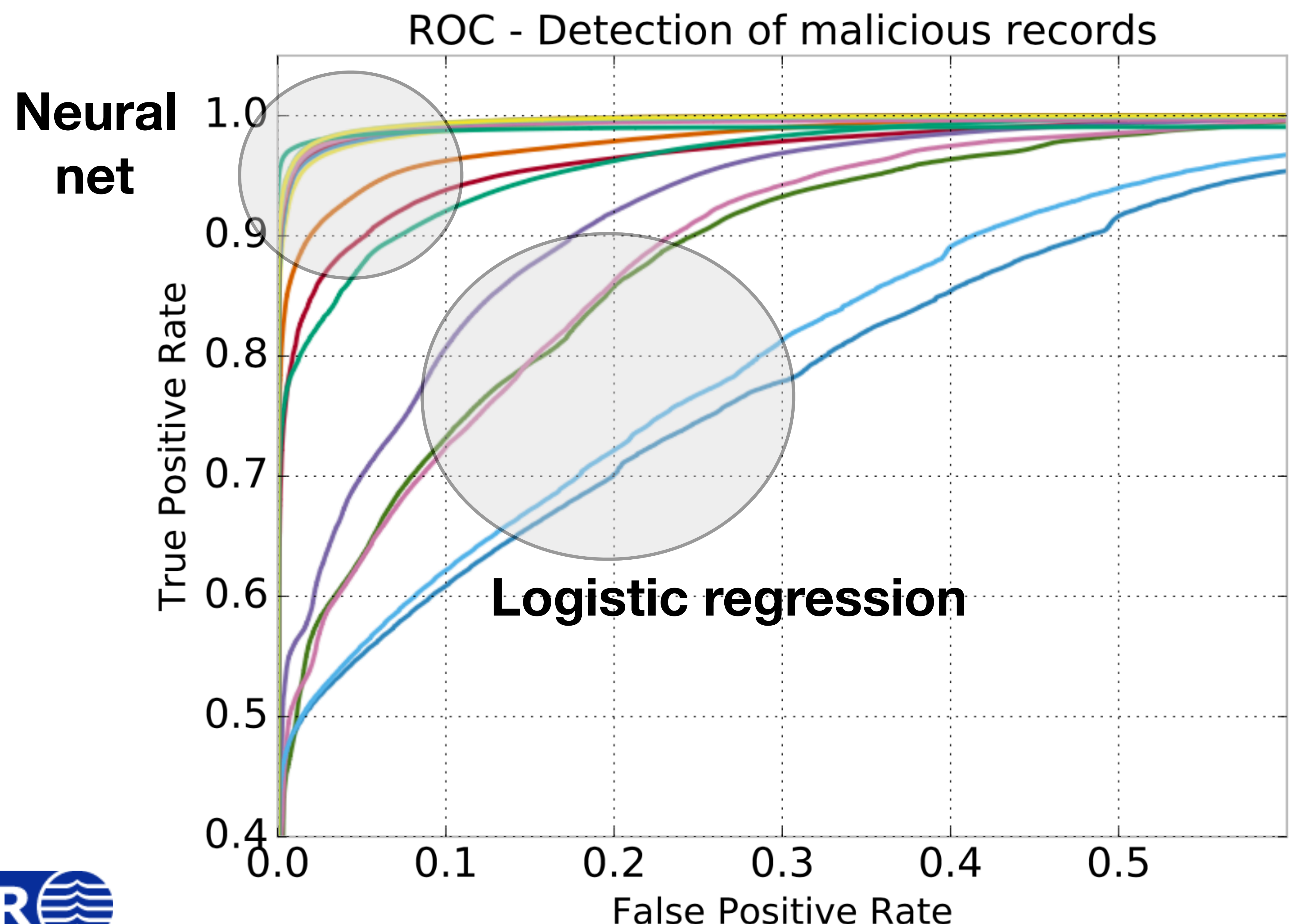


Results

Model	Benign P	Benign R	Malicious P	Malicious R	Accuracy
nb_domain_queries < 10	0.98	0.44	0.10	0.87	0.54
Logistic regression	0.97	0.97	0.60	0.65	0.944
Neural net (with 1 hidden layer)	0.99	0.99	0.93	0.93	0.990
Neural net (with 2 hidden layers)	1.00	0.99	0.92	0.95	0.990
Neural net (with 3 hidden layers and two passes)	1.00	**1.00**	0.97	0.96	**0.995**

ROC curve

[Lison, P. & Mavroeidis, V. (2017), Neural Reputation Models learned from Passive DNS Data. In *IEEE Big Data 2017*]



Conclusion

- Neural networks can be successfully used to predict the **reputation** of end-point hosts
  - Detection of DGA from the domain names
  - Detection of malicious records from passive DNS

- Can be integrated in software tools for cyber-threat intelligence

- Future work:
  - Integration of unstructured data sources (i.e. textual data)?