Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

# Neural models for predicting the reputation of end-point hosts

**Pierre Lison**

Norwegian Computing Center

AFSecurity

27.02.2018

# Introduction

- Blacklists and whitelists (= **reputation lists**) often employed to filter network traffic
- Shortcomings:
  - Complex, time-consuming (manual) process
  - Limited coverage
  - Static (can be circumvented through techniques such domain flux and fast flux networks)

# Outline

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

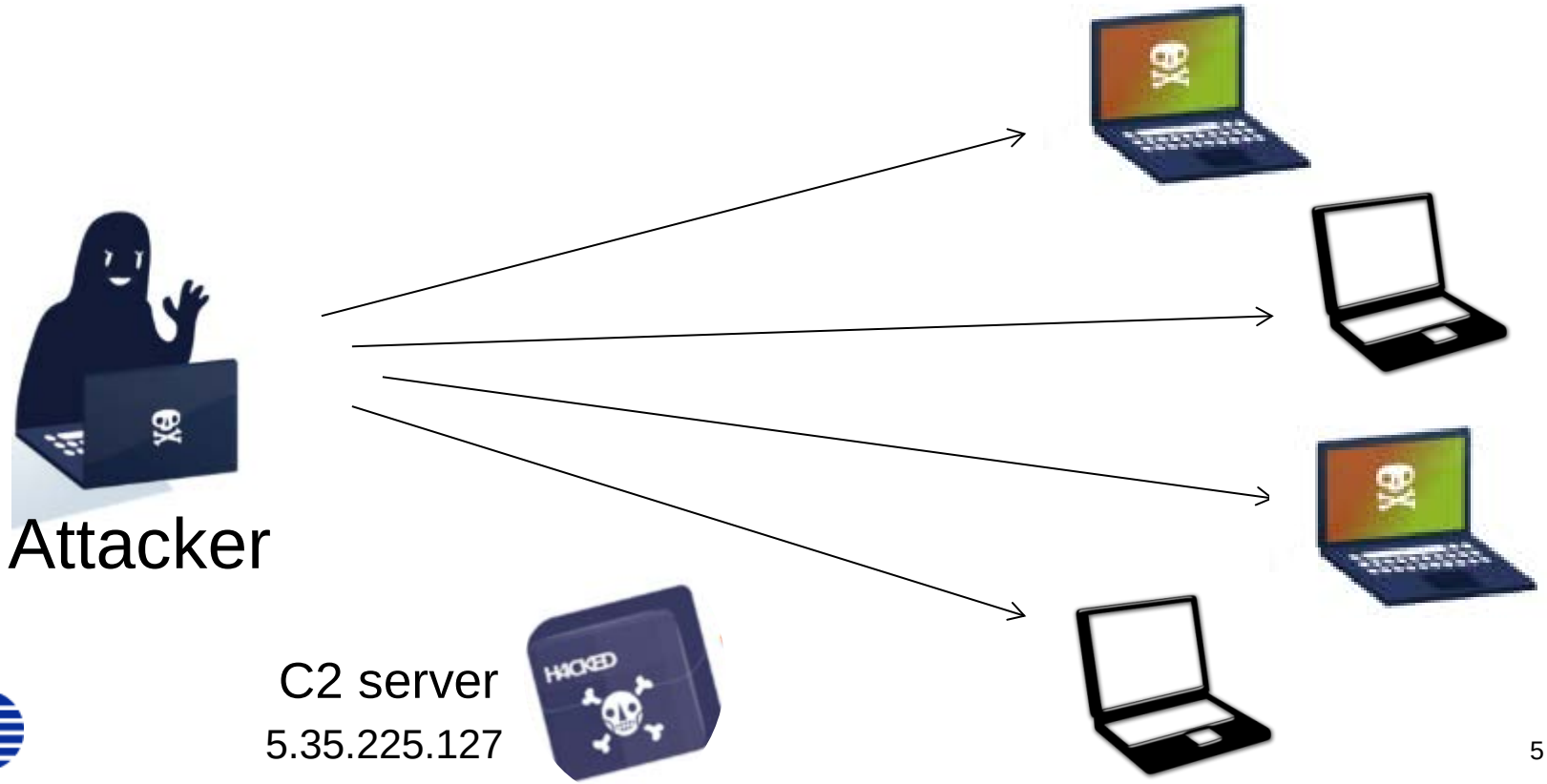
- Focus on two questions:
  1. Detecting malware-generated domain names through recurrent neural networks
  2. Predicting the reputation of domains and IP addresses from passive DNS data

# Outline

- Can we use machine learning to automatically predict the reputation of end-point hosts?
- Focus on two questions:
  1. **Detecting malware-generated domain names through recurrent neural networks**
  2. Predicting the reputation of domains and IP addresses from passive DNS data

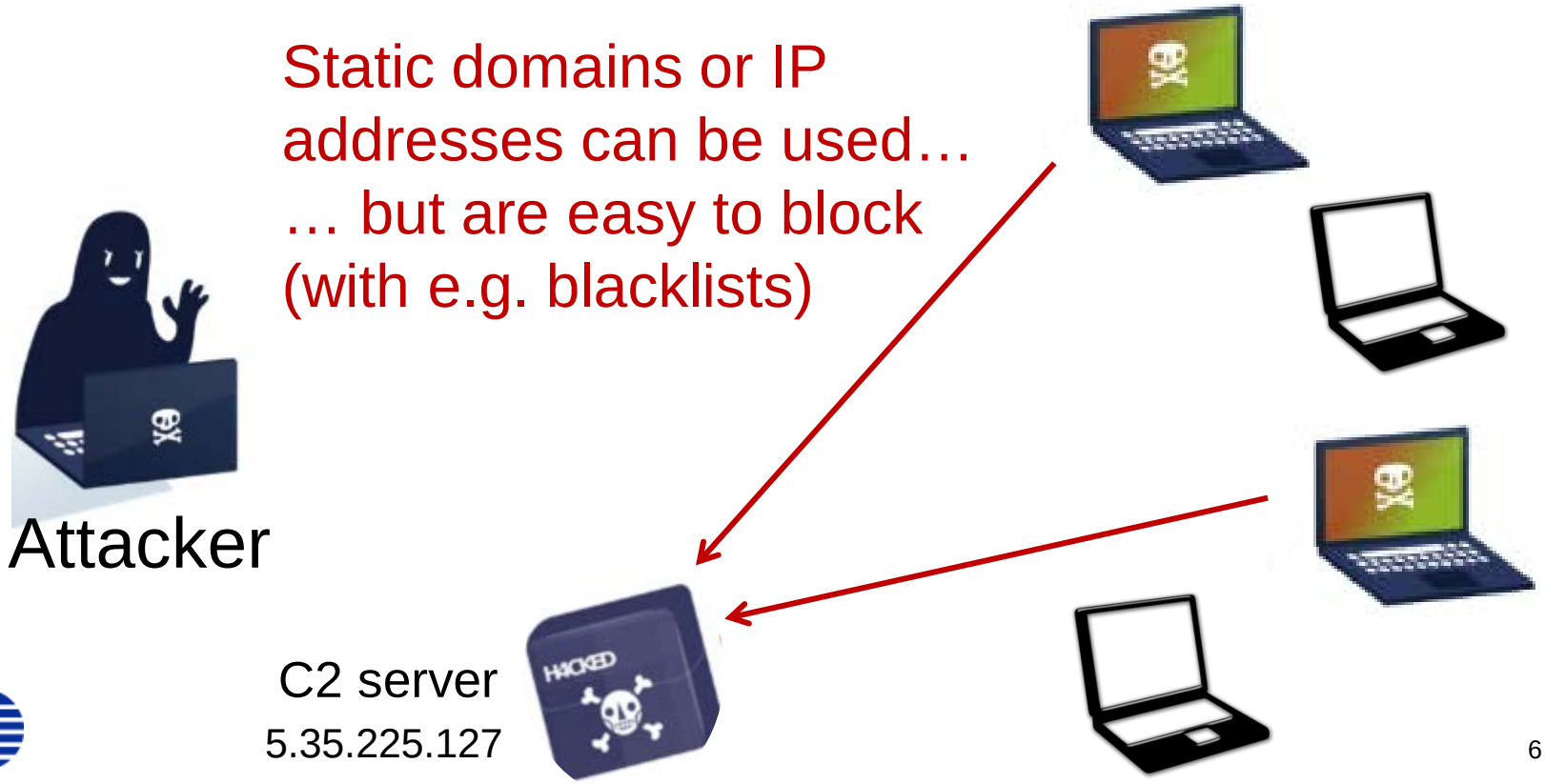
# Domain-generating malware

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



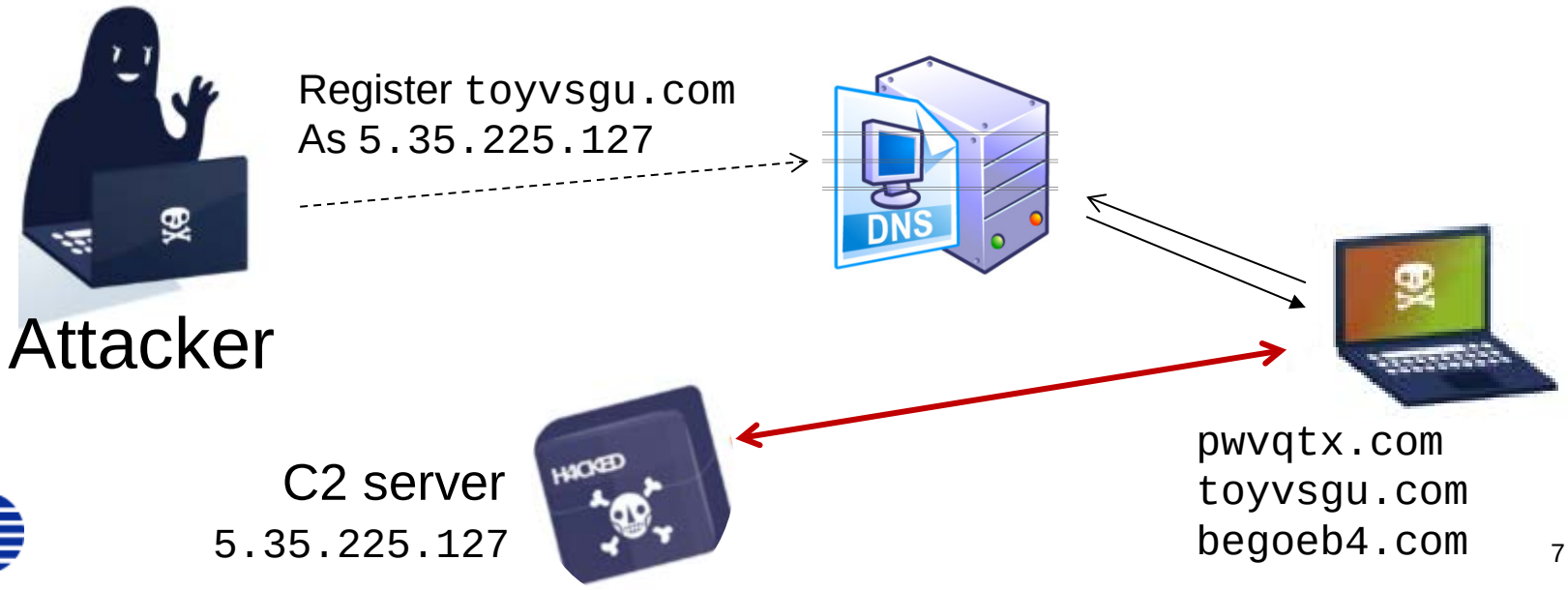
# Domain-generating malware

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



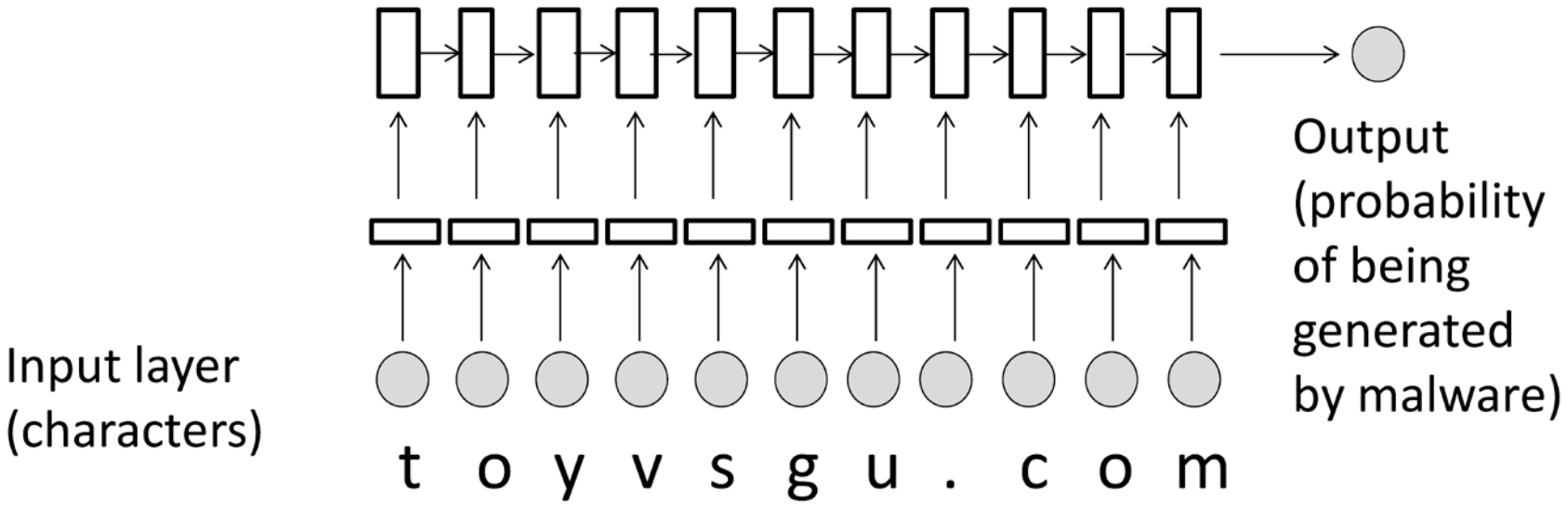
# Domain-generating malware

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

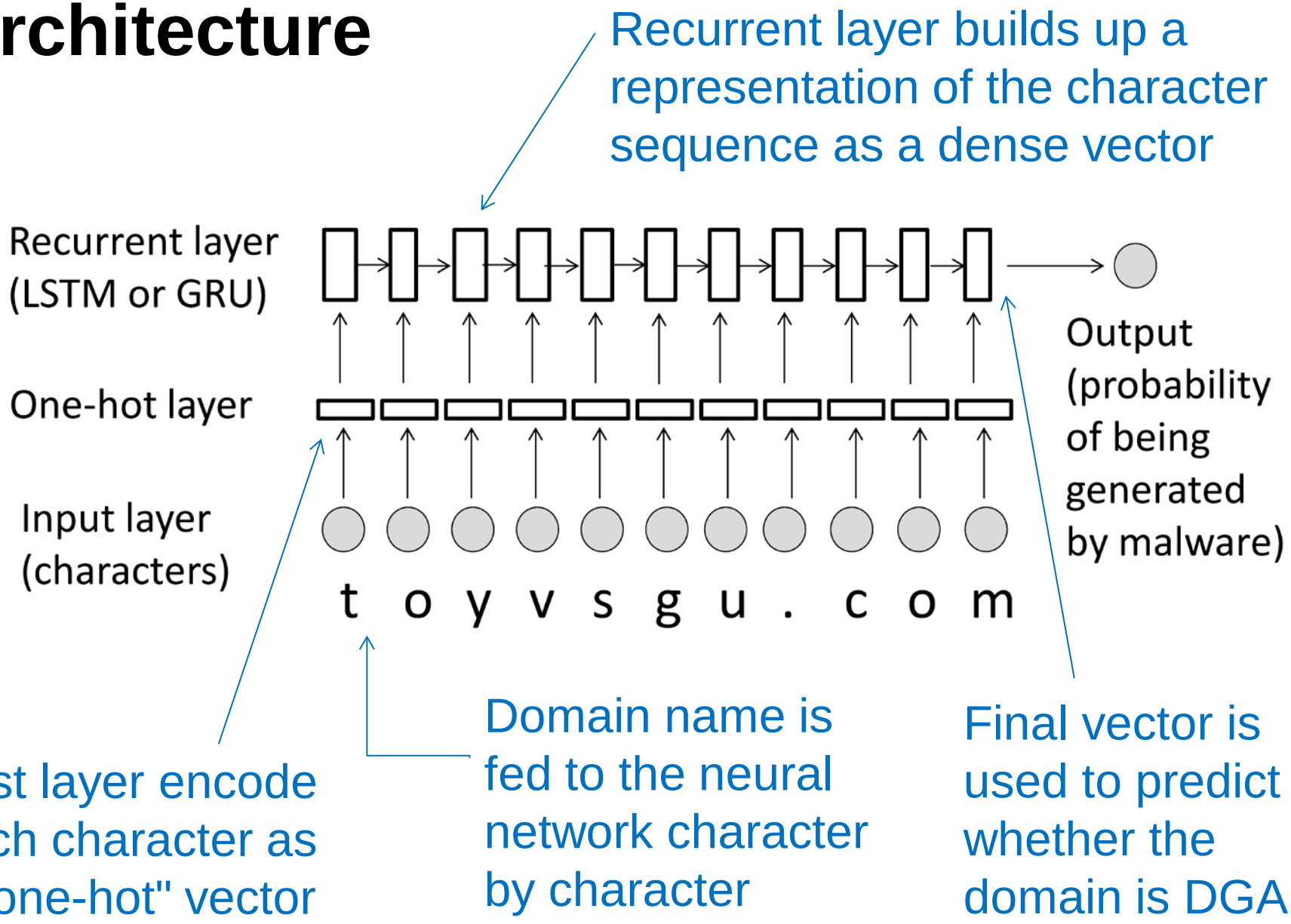


# Detection of DGAs

- **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



# Data

- The parameters of the neural model must be estimated from training data
- **Negative examples** (benign domains):
  - Snapshots from the Alexa top 1 million domains
  - Total: over 4 million domains
- **Positive examples** (malware DGAs)
  - DGA lists from the DGArchive (63 types of malware)
  - Feeds from Bambenek Consulting
  - Domain generators for 11 DGAs
  - Total: 2.9 million domains

# Data

| Malware | Frequency | Malware | Frequency | Malware | Frequency |
|---|---|---|---|---|---|
| bamital | 40 240 | gozi | 105 631 | ramdo | 15 984 |
| banjori | 89 984 | hesperbot | 370 | ramnit | 90 000 |
| bedep | 15 176 | locky | 179 204 | ranbyu | 40 000 |
| beebone | 420 | madmax | 192 | ranbyus | 12 720 |
| blackhole | 732 | matsnu | 12 714 | rovnix | 40 000 |
| bobax | 19 288 | modpack | 52 | shifu | 4 662 |
| conficker | 400 000 | murofet | 53 260 | simda | 38 421 |
| corebot | 50 240 | $\text{murofet}_w$ | 40 000 | sisron | 5 936 |
| cryptolocker | 55 984 | necur | 40 000 | suppobox | 41 014 |
| cryptowall | 94 | necurs | 36 864 | sutra | 9 882 |
| dircrypt | 11 110 | nymaim | 186 653 | symmi | 40 064 |
| dnschanger | 40 000 | oderoor | 3 833 | szribi | 16 007 |
| downloader | 60 | padcrypt | 35 616 | tempedreve | 453 |
| dyre | 47 998 | proslikefan | 75 270 | tinba | 80 000 |
| ekforward | 1 460 | pushdo | 176 770 | torpig | 40 000 |
| emotet | 40 576 | pushdotid | 6 000 | tsifiri | 59 |
| feodo | 192 | pykspa | 424 215 | urlzone | 34 536 |
| fobber | 2 600 | pykspa2 | 24 322 | vawtrak | 1 050 |
| gameover | 80 000 | qadars | 40 400 | virut | 400 600 |
| gameover_p2p | 41 000 | qakbot | 90 000 | volatilecedar | 1 494 |
| | | | | xxhex | 4400 |
| | | | | **Total** | 2 925 168 |

# Results

Area Under the Curve (AUC) of the ROC curve (see next slide)

- Detection

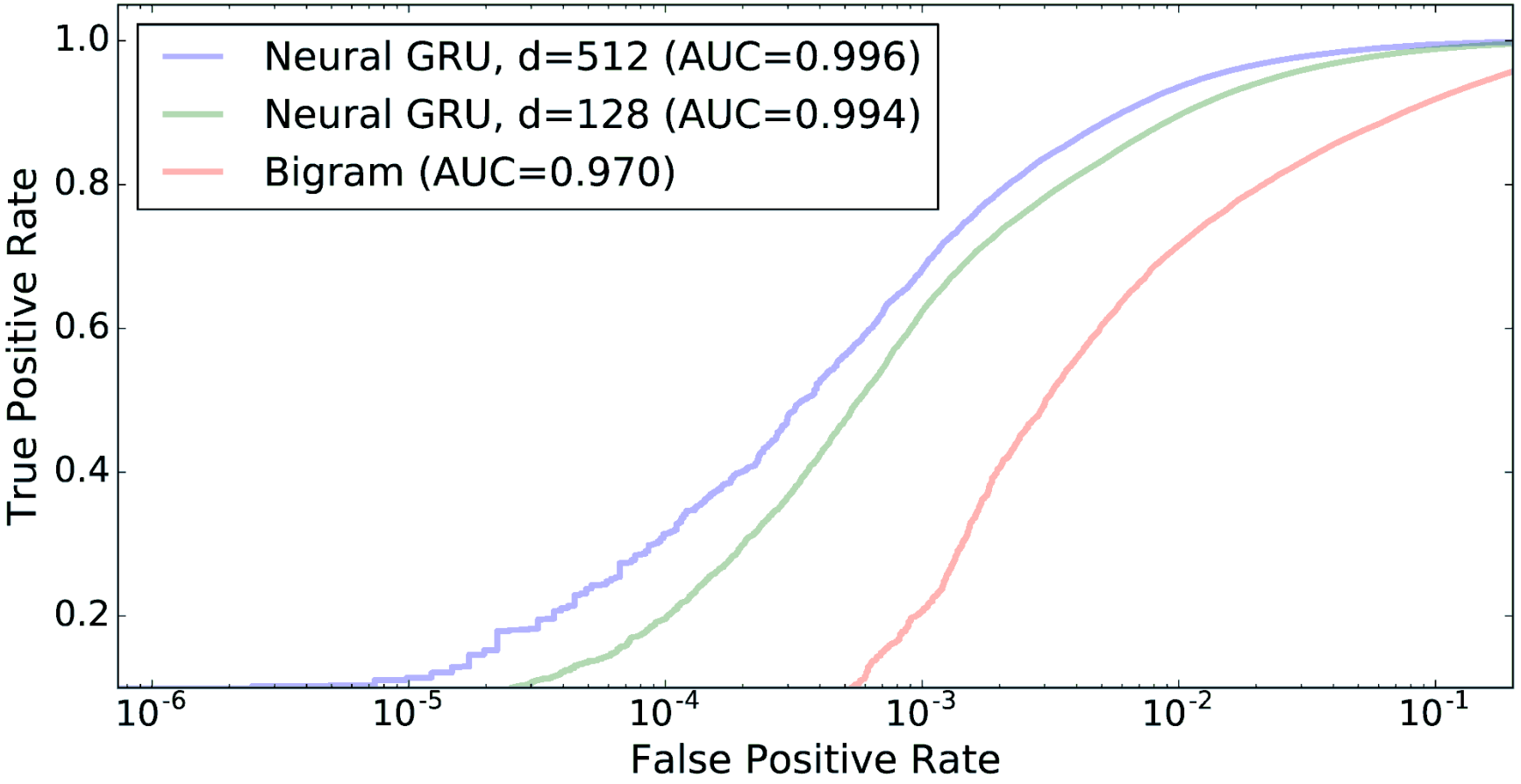
| | Accuracy | Precision | Recall | $F_1$ score | ROC AUC |
|---|---|---|---|---|---|
| Bigram | 0.915 | 0.927 | 0.882 | 0.904 | 0.970 |
| Neural model | **0.973** | **0.972** | **0.970** | **0.971** | **0.996** |

- Classification

| | Accuracy | Precision | | Recall | | $F_1$ score | |
|---|---|---|---|---|---|---|---|
| | | Micro | Macro | Micro | Macro | Micro | Macro |
| Bigram | 0.800 | 0.787 | 0.564 | 0.800 | 0.513 | 0.787 | 0.522 |
| Neural model | **0.892** | **0.891** | **0.713** | **0.892** | **0.653** | **0.887** | **0.660** |

Micro: weighted averages over all classes
Macro: unweighted averages

# ROC curve


Neural GRU, d=512 (AUC=0.996)
Neural GRU, d=128 (AUC=0.994)
Bigram (AUC=0.970)
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
$10^{-6}$
$10^{-5}$
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$

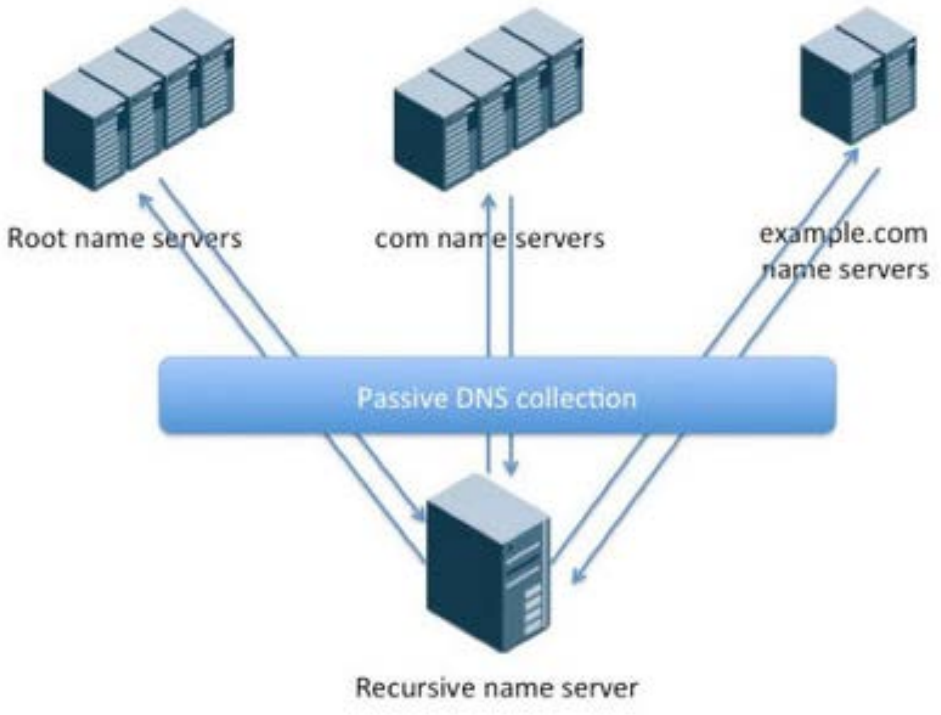
# Outline

- Can we use machine learning to automatically predict the reputation of end-point hosts?
- Focus on two questions:
  1. Detecting malware-generated domain names through recurrent neural networks
  2. **Predicting the reputation of domains and IP addresses from passive DNS data**

# Reputation models

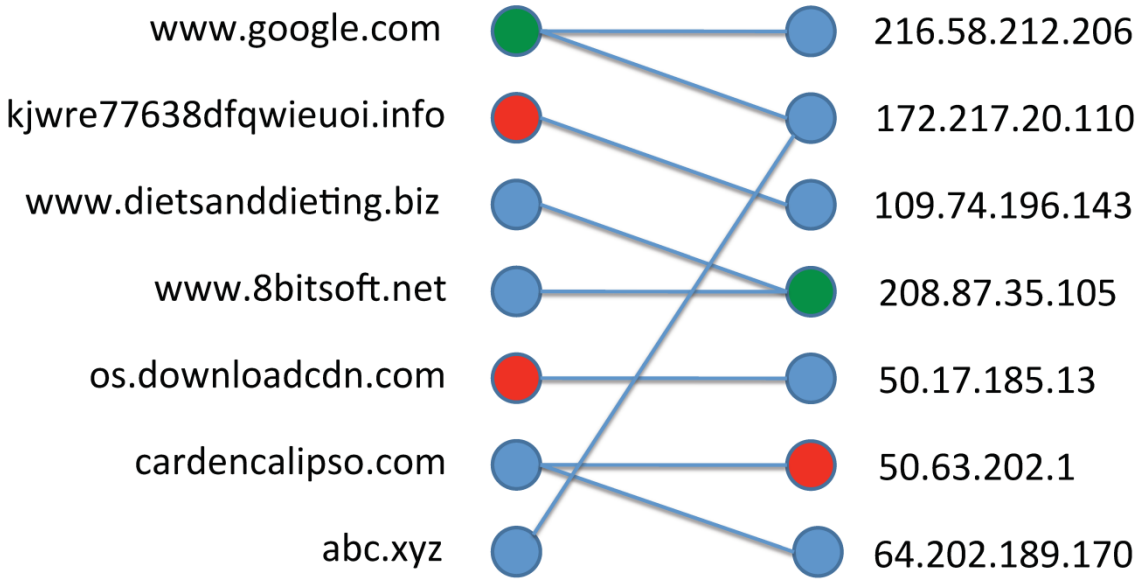
- Can we automatically predict the reputation of domain names and IP addresses from DNS data?

- Used a large passive DNS from Mnemonic:
  - *720 million* aggregated DNS queries collected over four years
  - *Server-to-server* communication (less privacy concerns)



# 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)



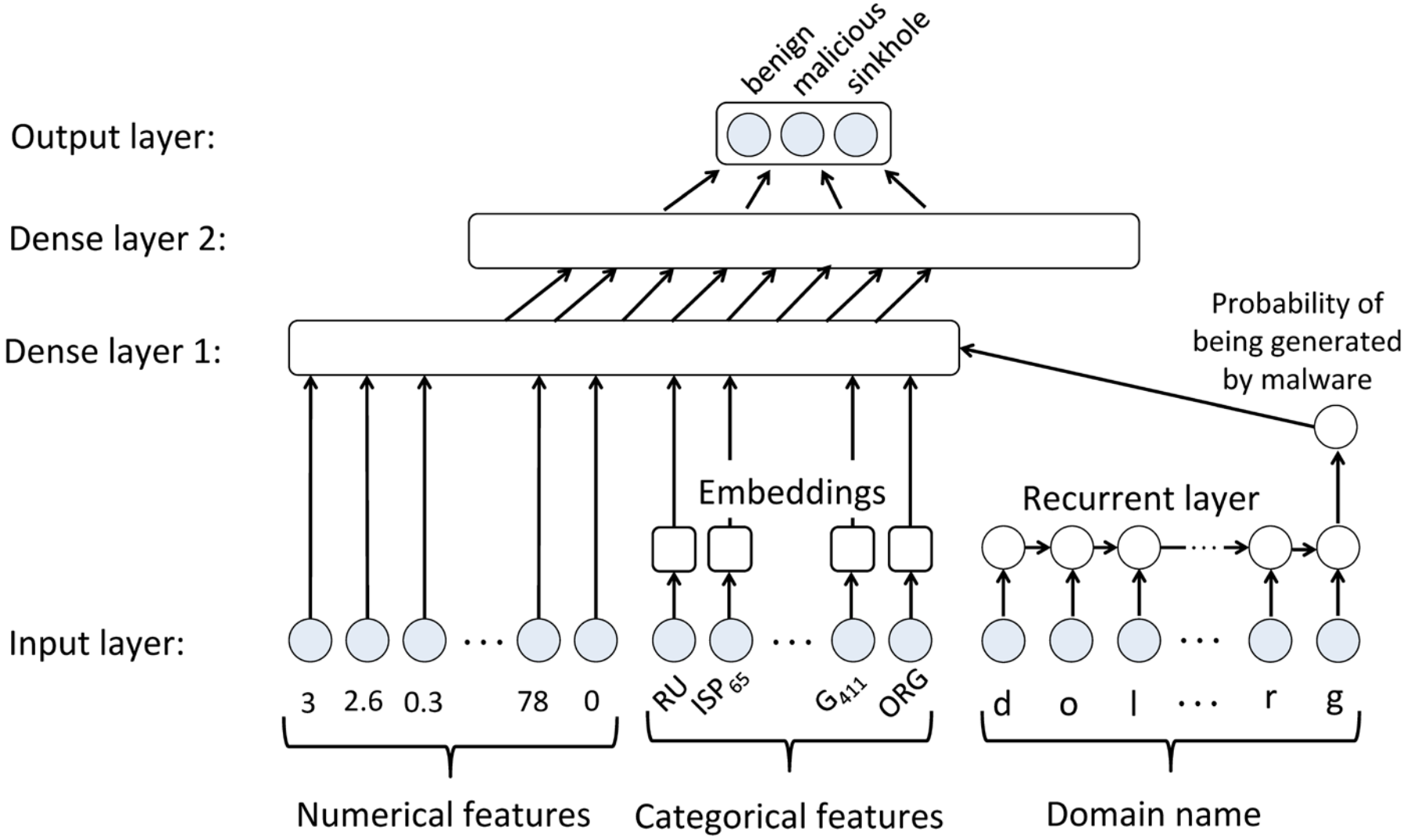

We enriched the passive DNS data with:

- 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 neighbouring records
  - Number of records at distance 1 and of reputation X
- Sequence of characters from the domain

# Neural model



# Results

| Model | Benign | | | Malicious | | | Sinkhole | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | |
| nb_domain_queries $< 10$ | 0.98 | 0.44 | 0.61 | 0.10 | 0.87 | 0.19 | 0.0 | 0.0 | 0.0 | 0.54 |
| Logistic regression | 0.97 | 0.97 | 0.97 | 0.60 | 0.65 | 0.62 | 0.51 | 0.26 | 0.35 | 0.944 |
| Neural net (with 1 hidden layer) | 0.99 | 0.99 | 0.99 | 0.93 | 0.93 | 0.93 | 0.99 | 1.00 | 0.99 | 0.990 |
| Neural net (with 2 hidden layers) | 1.00 | 0.99 | 0.99 | 0.92 | 0.95 | 0.93 | 0.98 | 1.00 | 0.99 | 0.990 |
| Neural net (with 3 hidden layers and two passes) | 1.00 | **1.00** | 1.00 | 0.97 | 0.96 | **0.96** | 0.99 | 0.96 | 0.98 | **0.995** |

In this setting, the neural net is first trained on the labelled dataset and applied to predict the reputation of unlabelled records, which are then used to get better estimates of the "neighbour" features. The model is then trained again on these new feature values.

# 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
- Current work:
  - Consolidate experimental results
  - Integration of unstructured data sources (i.e. textual data)