# FANCI

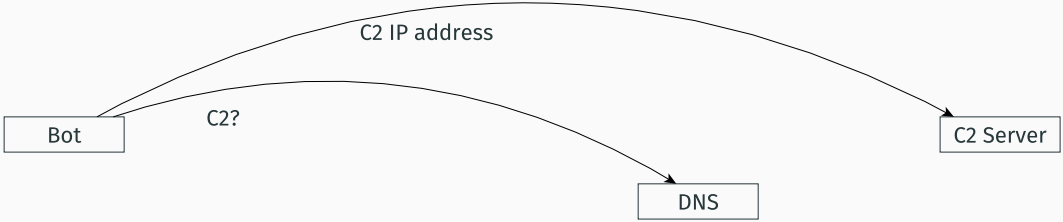
## Feature-based Automated NXDomain Classification and Intelligence

Samuel Schüppen[1] Dominik Teubert[2] Patrick Herrmann[1] Ulrike Meyer[1]

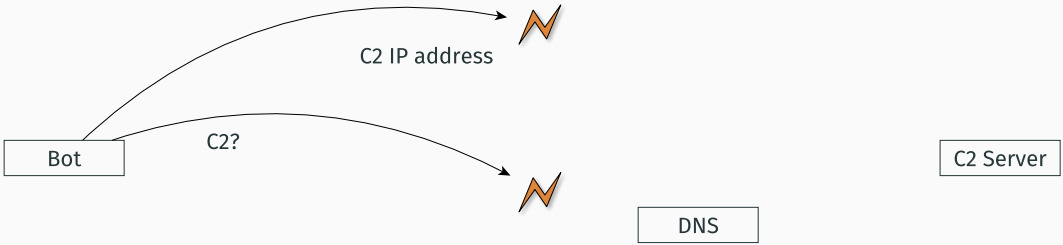
August 17, 2018

RWTH Aachen University[1] and Siemens CERT[2]

# Traditional Bot Communication

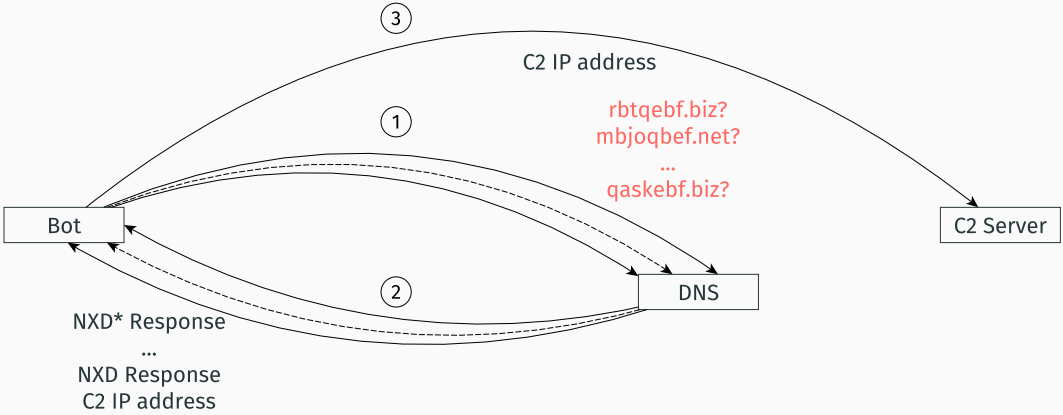


# Traditional Bot Communication




Easy to block fixed domains or IP addresses.

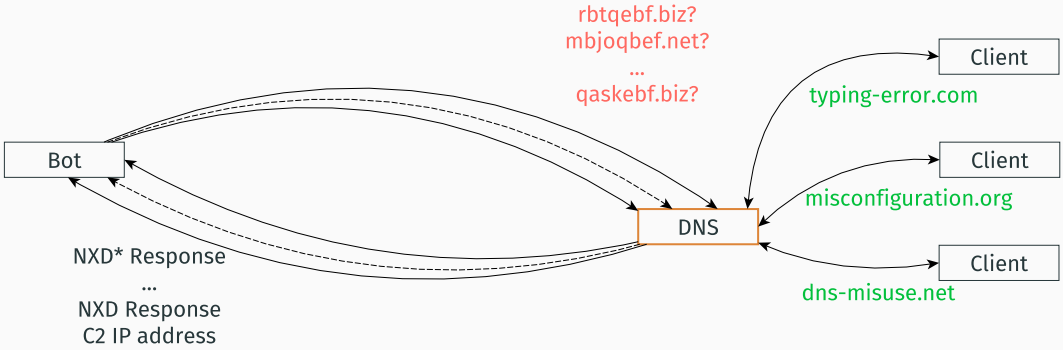
# Domain Generation Algorithm (DGA)




*non-existent domain (NXD)

DGAs are popular: more than 70 DGAs known

# The Approach




*non-existent domain (NXD)

## Classify non-existent domains (NXDs)

iieee.org
i8a0q2wdu8otulkfylo2gdq.ddns.net
mcirosfot.com
layergetadobpeflash.net
strangerbiketechnology.com
etadobeflgashplayer.net
bayanescortbandirma.xyz
fsztakqwdjfqsc.asa.at
wfnfhde.de
kaqoeizerbo
ahxurofbdughh.rwth-aachen.de
873c174ca173b5393e93f9571e8a293b.org
de-swyx-2.fraba.local
kh1her76avy0qnelivijwd1.ddns.net
www.digitex-eu.com
brwc0f8da79205c.net

# The Challenge

## Classify non-existent domains (NXDs)

iieee.org
i8a0q2wdu8otulkfylo2gdq.ddns.net
mcirosfot.com
layergetadobpeflash.net
strangerbiketechnology.com
etadobeflgashplayer.net
bayanescortbandirma.xyz
fsztakqwdjfqsc.asa.at

wfnfhde.de
kaqoeizerbo
ahxurofbdughh.rwth-aachen.de
873c174ca173b5393e93f9571e8a293b.org
de-swyx-2.fraba.local
kh1her76avy0qnelivijwd1.ddns.net
www.digitex-eu.com
brwc0f8da79205c.net

# Contribution

## We present FANCI

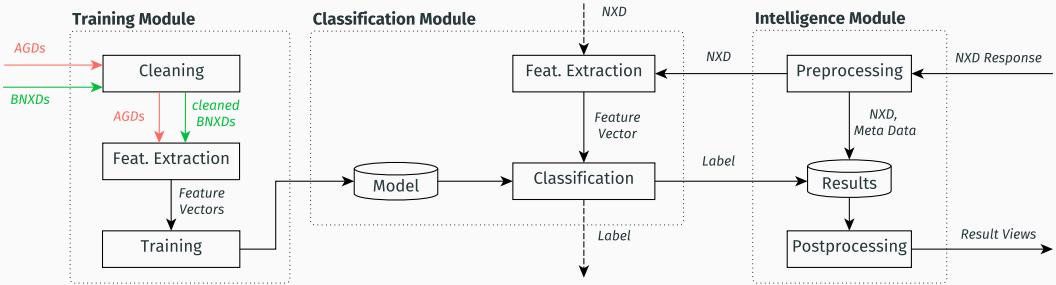
Feature-based Automated NXDomain Classification Intelligence

ML-based classification of NXDs into benign and DGA-related
High accuracy
Lightweight and efficient
Classification on domain names only — no meta data required
Generalizable and usable as-a-service

Find DGA-based malware infected devices / deliver actionable IoCs

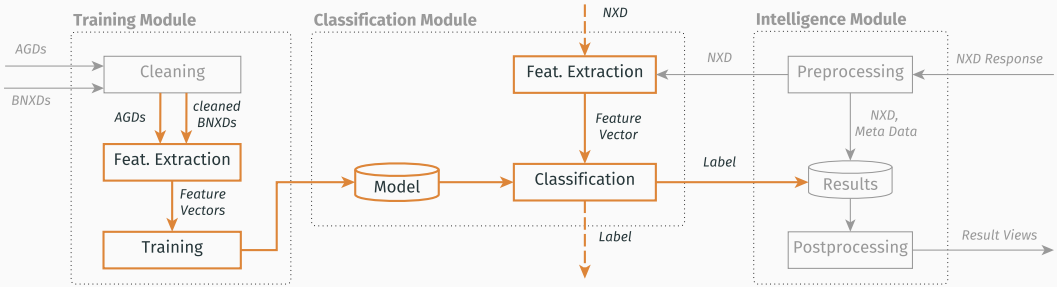
# FANCI

# FANCI's Architecture




benign non-existent domain (BNXD)
algorithmically generated domain (AGD)

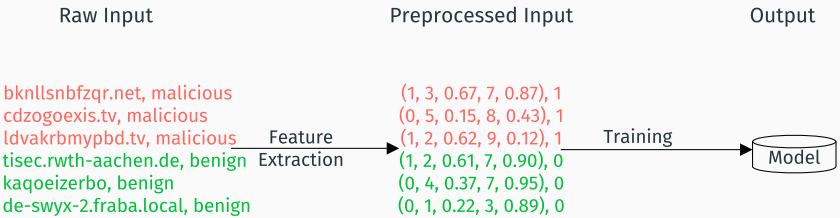
# Classification



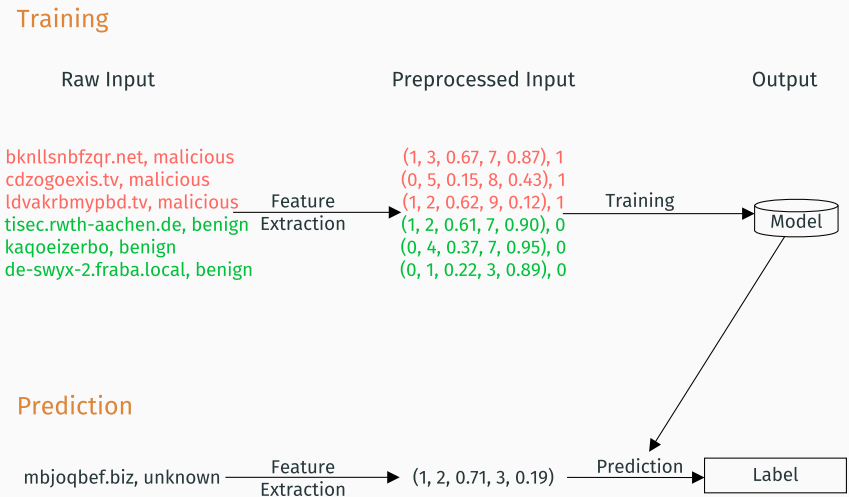

benign non-existent domain (BNXD)
algorithmically generated domain (AGD)

# Supervised Learning Classifier

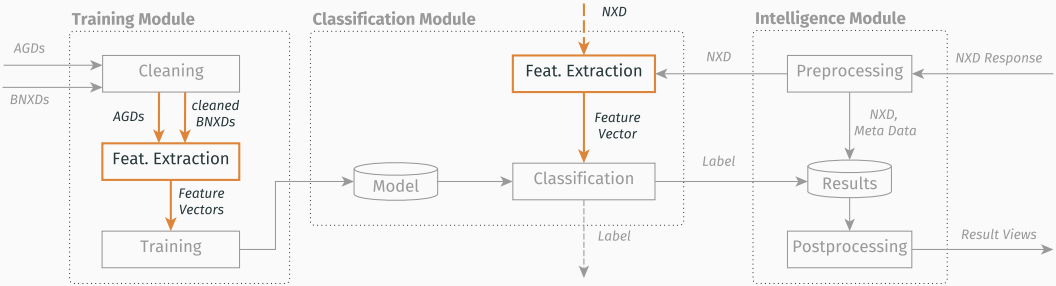
## Training



# Supervised Learning Classifier



# Feature Extraction



# Feature Design


| 3 Categories | 21 overall | Example | Output | $\mathcal{F}(d_1)$ | $\mathcal{F}(d_2)$ |
|---|---|---|---|---|---|
| structural | 12 | Valid TLD | binary | 1 | 1 |
| linguistic | 7 | Alphabet Cardinality | integer | 10 | 21 |
| statistical | 2 | Entropy | rational | 3.35 | 4.18 |

$d_1 =$ `nxdomain.siemens.com`
$d_2 =$ `dekh1her76avy0qnelivijwd1.ddns.net`

# Evaluation

# Data Set

RWTH Aachen University

- 35.8 million unique benign NXDs
- 31 days (May - June '17)
- central DNS server

Siemens AG

- 31.2 million unique benign NXDs
- 61 days (Sep - Oct '17)
- DNS servers from EU, USA, Asia

DGArchive[1]

- 49.7 million unique DGA domains
- 1,344 days (2014 - 2018)
- more than 70 DGAs

[1]Thanks a lot to Daniel Plohmann and Fraunhofer for granting us access to DGArchive `dgarchive.caad.fkie.fraunhofer.de`

# Select the Best Classifier

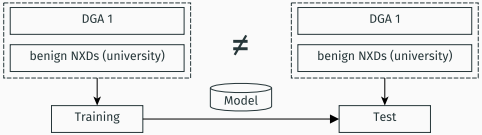
One DGA one Classifier vs. Multiple DGAs one Classifier

---

Support Vector Machine vs. Random Forest

# Select the Best Classifier

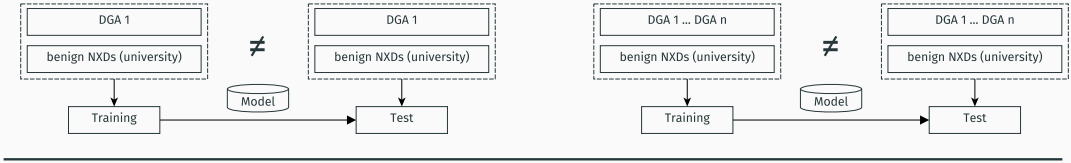
One DGA one Classifier vs. Multiple DGAs one Classifier




Support Vector Machine vs. Random Forest

# Select the Best Classifier

One DGA one Classifier vs. Multiple DGAs one Classifier



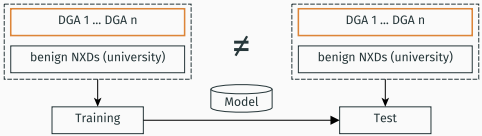

Support Vector Machine vs. Random Forest

# The Best Performing Classifier

One Random Forest classifier trained with all known DGAs

# Multiple DGAs One Random Forest Classifier

20 data sets (size 100,000), 5 repetitions per 5-fold cross-validation: $20 \cdot 5 \cdot 5 = 500$ passes

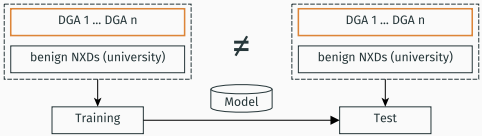



| RF | $\bar{x}$ | $\sigma$ | $x_{min}$ | $\tilde{x}$ | $x_{max}$ |
|---|---|---|---|---|---|
| ACC | 0.9976 | 0.0001 | 0.9975 | 0.9976 | 0.9978 |
| TPR | 0.9976 | 0.0001 | 0.9974 | 0.9976 | 0.9978 |
| FPR | 0.0025 | 0.0001 | 0.0023 | 0.0025 | 0.0027 |

$\bar{x}$: arithmetic mean; $\sigma$: standard deviation; $x_{min}$: minimum; $\tilde{x}$: median; $x_{max}$: maximum

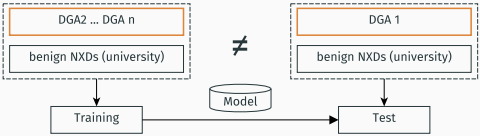
# Multiple DGAs One Random Forest Classifier

20 data sets (size 100,000), 5 repetitions per 5-fold cross-validation: $20 \cdot 5 \cdot 5 = 500$ passes




| RF | $\bar{x}$ | $\sigma$ | $x_{min}$ | $\tilde{x}$ | $x_{max}$ |
|---|---|---|---|---|---|
| ACC | 0.9976 | 0.0001 | 0.9975 | 0.9976 | 0.9978 |
| TPR | 0.9976 | 0.0001 | 0.9974 | 0.9976 | 0.9978 |
| FPR | 0.0025 | 0.0001 | 0.0023 | 0.0025 | 0.0027 |

20 data sets (size 100,000), leave-one-out: $20 \cdot 59 = 1180$

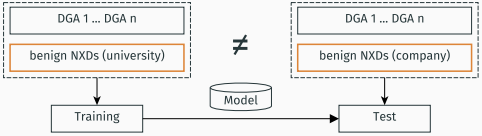


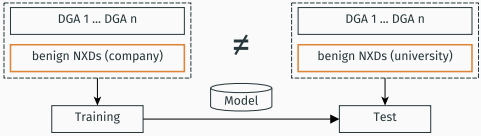

| RF | $\bar{x}$ | $\sigma$ | $x_{min}$ | $\tilde{x}$ | $x_{max}$ |
|---|---|---|---|---|---|
| ACC | 0.9807 | 0.0003 | 0.9797 | 0.9808 | 0.9812 |
| TPR | 0.9639 | 0.0007 | 0.9618 | 0.9640 | 0.9647 |
| FPR | 0.0024 | 0.0002 | 0.0022 | 0.0024 | 0.0027 |

$\bar{x}$: arithmetic mean; $\sigma$: standard deviation; $x_{min}$: minimum; $\tilde{x}$: median; $x_{max}$: maximum

# Generalization

20 data sets (size 100,000) for each benign data source: $20 \cdot 20 = 400$ passes



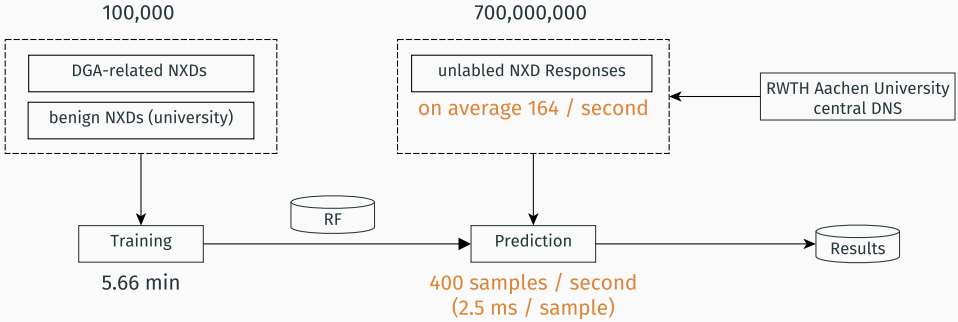

| RF | $\bar{x}$ | $\sigma$ | $x_{min}$ | $\tilde{x}$ | $x_{max}$ |
|---|---|---|---|---|---|
| ACC | 0.9953 | 0.0002 | 0.9951 | 0.9953 | 0.9957 |
| TPR | 0.9994 | 0.0001 | 0.9992 | 0.9994 | 0.9995 |
| FPR | 0.0087 | 0.0003 | 0.0080 | 0.0088 | 0.0092 |

| RF | $\bar{x}$ | $\sigma$ | $x_{min}$ | $\tilde{x}$ | $x_{max}$ |
|---|---|---|---|---|---|
| ACC | 0.9979 | 0.0001 | 0.9977 | 0.9978 | 0.9980 |
| TPR | 0.9995 | 0.0001 | 0.9994 | 0.9995 | 0.9995 |
| FPR | 0.0038 | 0.0002 | 0.0035 | 0.0038 | 0.0041 |

$\bar{x}$: arithmetic mean; $\sigma$: standard deviation; $x_{min}$: minimum; $\tilde{x}$: median; $x_{max}$: maximum

# Real World Application: Setup

- Fresh data set (university): 31 days (13 Oct ’17 - 12 Nov ’17)
- 35 million unique NXDs
- Single-threaded on consumer hardware



# Real World Application: Results

We found 10 groups of unknown DGA-related domains: 405 overall

### Unknown DGA 1

```
cxoriilg.host
dcveyroohhuz.host
ktnotybgnqrjvkq.host
ndptbhn.host
qbeweonxhzlflh.host
zwchzomnkersegz.host
```

### Unknown DGA 2

```
blwemxb.ga
yinnic.gq
fyrrzx.ml
fhvfbhq.tk
ihrslrk.cf
xlajbu.cf
```

### Unknown DGA 3

```
brn001ba9933850.net
brn001ba99fa1c7.net
brw48e244240e9d.net
brwc0f8da79205c.net
brw184f328b61dc.net
brwc48e8fbdfa3e.net
```

### Unknown DGA 4 or Seed of *Redyms*

```
3fdqrbnum3fa2j1.3tfrmn27i.com
c4xf33p7nrvo3l0h.23bjj3a0.com
wpdcp7uym0.up18xtxzouumzd.com
wlh8tj55fxfh.n51ah7y227y-.com
0qpumb2ks982r6.ytt2ot0foi.com
xgs66mu-uig2u.cjswb3q4m45.com
```

### Unknown DGA 5 or Seed of *GozNym 2nd Stage | Nymaim*

```
afyonescortkizlar.xyz
ordubayanescort.xyz
kirikkalebayanescort.xyz
nigdebayanescort.xyz
bayanescortbandirma.xyz
bayanescortbilecik.xyz
afyonescortkizlar.xyz
```

### Unknown DGA 6 or Seed of *GozNym 2nd Stage | Nymaim*

```
getbeautifuljacked.xyz
evelynmiller.xyz
juicepress.xyz
quietbranch.xyz
tracyhernandez.xyz
webhostpremium.xyz
wertvollebrillanthobby.xyz
```

# Summary

## FANCI classifies NXDs into benign and DGA-related

- based on domain names only
- lightweight, efficient, and with high accuracy

## FANCI is real-world applicable

- found 10 groups of unknown DGA domains

## FANCI generalizes well and is usable as-a-service

Available on GitHub: `https://github.com/fanci-dga-detection/fanci`