

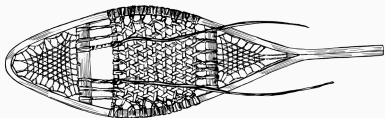
Melting the Snow

Using Active DNS Measurements to Detect Snowshoe Spam Domains

Olivier van der Toorn

February 1, 2018

University of Twente, Design and Analysis of Communication Systems



Introduction



[matrix]

@lordievader:corellian.student.utwente.nl



o.i.vandertoorn@utwente.nl

HELLO

My name is

Olivier

Introduction



HELLO

My name is

Olivier

Introduction



Introduction

HELLO

My name is

Olivier

We hypothesize that the use of active DNS measurements is a good way to detect snowshoe spam domains.



How can we detect snowshoe spam domains through the use of active DNS measurements?



Introduction

HELLO

My name is

Olivier

Are we able to automate the detection
of snowshoe spam domains?



Introduction

HELLO

My name is

Olivier

What are the advantages of this approach over other approaches?



Introduction

HELLO

My name is

Olivier

What are the advantages of this approach over other approaches?
How large is the time advantage of this approach?

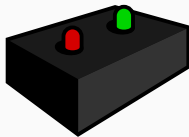




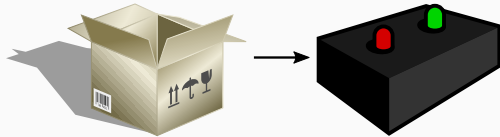
What are the advantages of this approach over other approaches?
How large is the time advantage of this approach?
How much more spam is blocked because of this method?

Overview



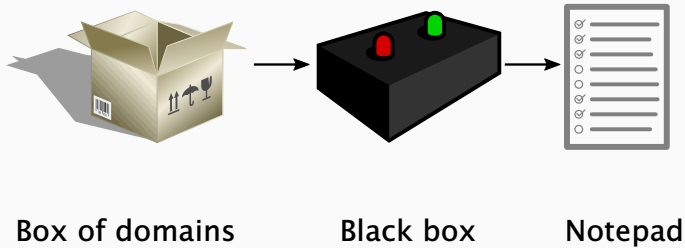


Black box



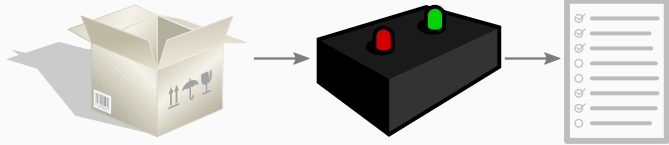
Box of domains

Black box



A Closer Look

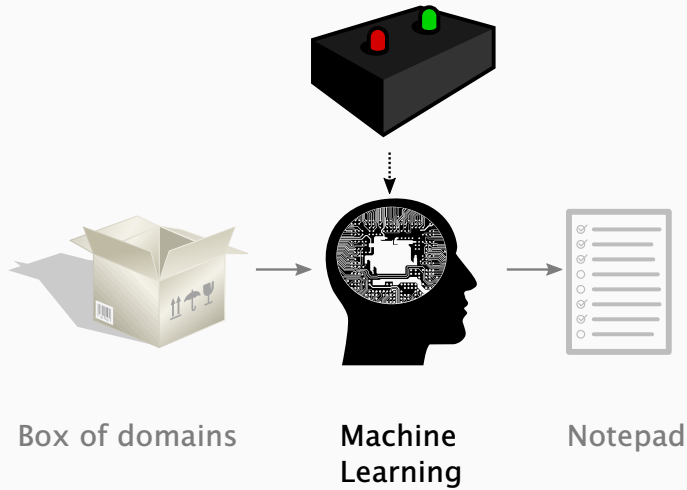


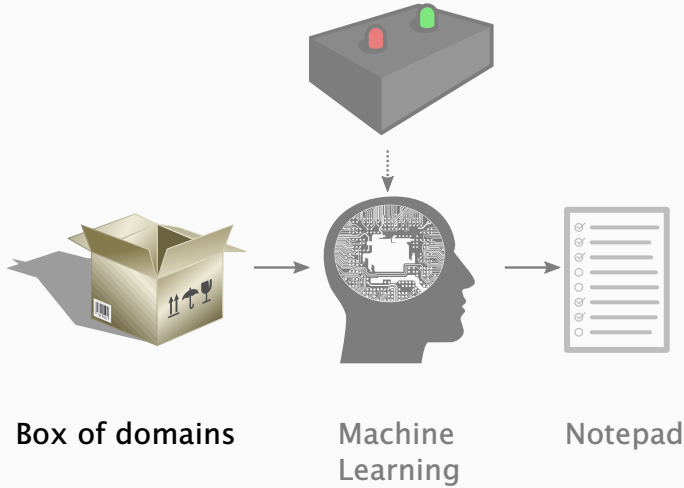


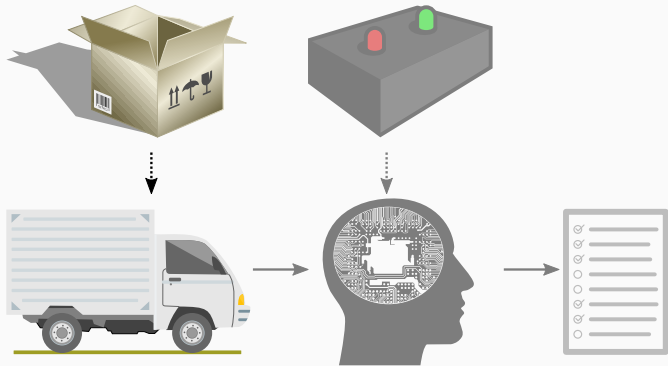
Box of domains

Black box

Notepad



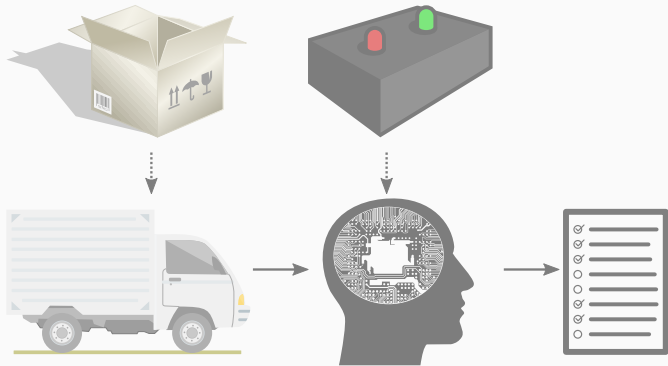




OpenINTEL

Machine
Learning

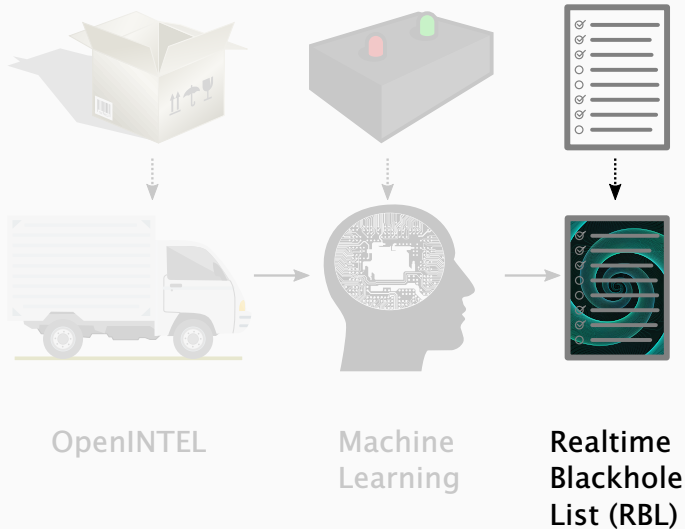
Notepad

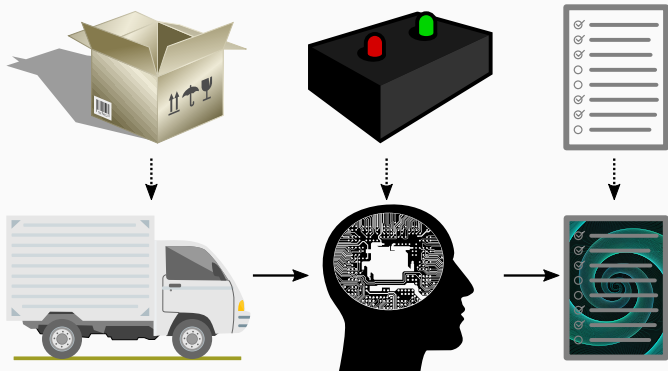


OpenINTEL

Machine
Learning

Notepad

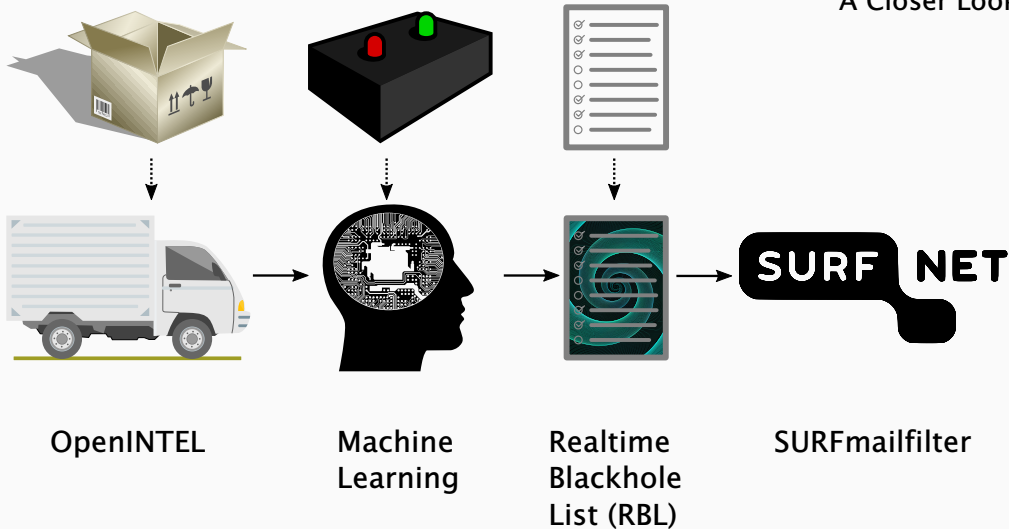




OpenINTEL

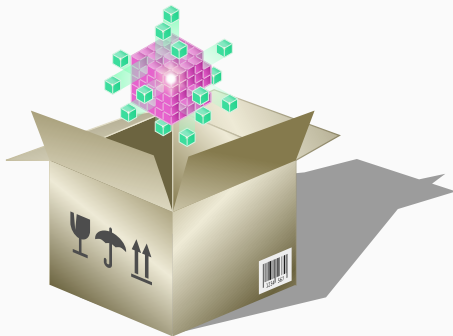
Machine
Learning

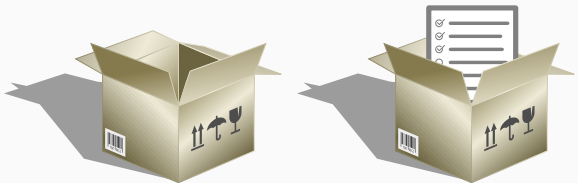
Realtime
Blackhole
List (RBL)



OpenINTEL





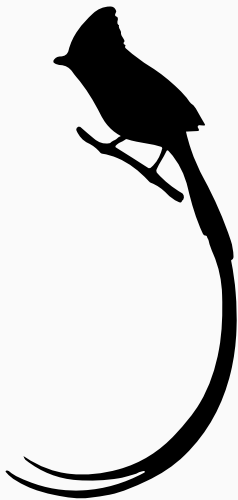


A dataset

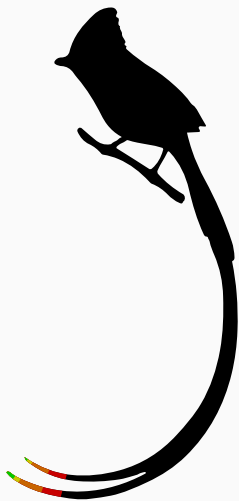
A labeled dataset



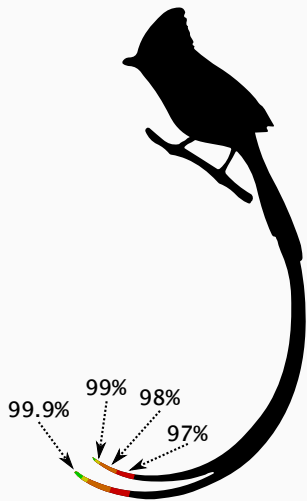
OpenINTEL



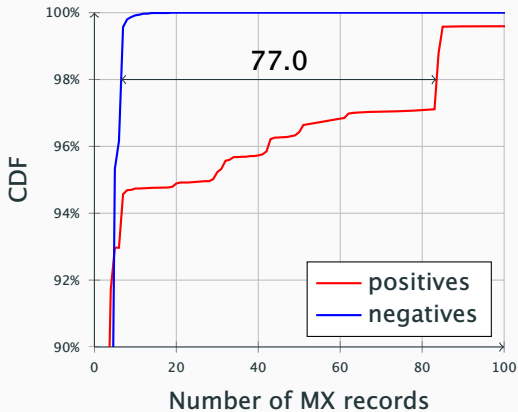
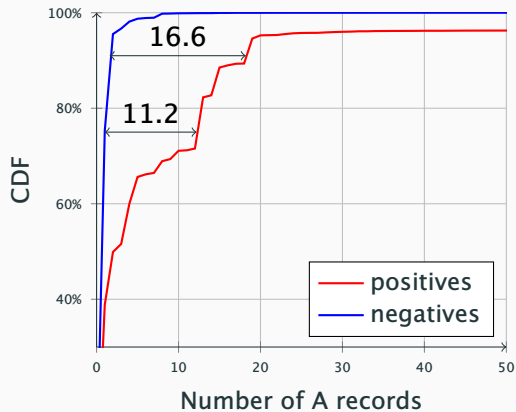
The long tail of the DNS



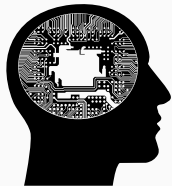
The **long** tail of the DNS



The **long** tail of the DNS



Machine Learning







	Spam		Ham	
Type	TP	FN	FP	TN



Type	Spam		Ham	
	TP	FN	FP	TN
BernoulliNB	12995	1535	2507	8344
GradientBoostingClassifier	12645	1885	9605	1246
MultinomialNB	12179	2351	1397	9454
RandomForestClassifier	11156	3374	1488	9363
MLPClassifier	7273	7257	707	10144
DecisionTreeClassifier	6279	8251	695	10156
AdaBoostClassifier	5971	8559	164	10687
KNeighborsClassifier	4562	9968	676	10175
SGDClassifier	3599	10931	674	10177



Type	Spam		Ham	
	TP	FN	FP	TN
RadiusNeighborsClassifier	13318	1212	2367	8484
BernoulliNB	12995	1535	2507	8344
GradientBoostingClassifier	12645	1885	9605	1246
MultinomialNB	12179	2351	1397	9454
RandomForestClassifier	11156	3374	1488	9363
MLPClassifier	7273	7257	707	10144
DecisionTreeClassifier	6279	8251	695	10156
AdaBoostClassifier	5971	8559	164	10687
KNeighborsClassifier	4562	9968	676	10175
SGDClassifier	3599	10931	674	10177



Type	Spam		Ham	
	TP	FN	FP	TN
GaussianNB	13330	1200	2075	8776
RadiusNeighborsClassifier	13318	1212	2367	8484
BernoulliNB	12995	1535	2507	8344
GradientBoostingClassifier	12645	1885	9605	1246
MultinomialNB	12179	2351	1397	9454
RandomForestClassifier	11156	3374	1488	9363
MLPClassifier	7273	7257	707	10144
DecisionTreeClassifier	6279	8251	695	10156
AdaBoostClassifier	5971	8559	164	10687
KNeighborsClassifier	4562	9968	676	10175
SGDClassifier	3599	10931	674	10177



Type	Spam		Ham	
	TP	FN	FP	TN
SVC	13449	1081	2339	8512
GaussianNB	13330	1200	2075	8776
RadiusNeighborsClassifier	13318	1212	2367	8484
BernoulliNB	12995	1535	2507	8344
GradientBoostingClassifier	12645	1885	9605	1246
MultinomialNB	12179	2351	1397	9454
RandomForestClassifier	11156	3374	1488	9363
MLPClassifier	7273	7257	707	10144
DecisionTreeClassifier	6279	8251	695	10156
AdaBoostClassifier	5971	8559	164	10687
KNeighborsClassifier	4562	9968	676	10175
SGDClassifier	3599	10931	674	10177



$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$



	Spam		Ham		
Type	TP	FN	FP	TN	Precision
SVC	13449	1081	2339	8512	85.18%
GaussianNB	13330	1200	2075	8776	86.53%
RadiusNeighborsClassifier	13318	1212	2367	8484	84.90%
BernoulliNB	12995	1535	2507	8344	83.82%
GradientBoostingClassifier	12645	1885	9605	1246	56.83%
MultinomialNB	12179	2351	1397	9454	89.70%
RandomForestClassifier	11156	3374	1488	9363	88.23%
MLPClassifier	7273	7257	707	10144	91.14%
DecisionTreeClassifier	6279	8251	695	10156	90.03%
AdaBoostClassifier	5971	8559	164	10687	97.32%
KNeighborsClassifier	4562	9968	676	10175	87.09%
SGDClassifier	3599	10931	674	10177	84.22%



Type	Spam		Ham		Precision
	TP	FN	FP	TN	
AdaBoostClassifier	5971	8559	164	10687	97.32%
MLPClassifier	7273	7257	707	10144	91.14%
DecisionTreeClassifier	6279	8251	695	10156	90.03%
MultinomialNB	12179	2351	1397	9454	89.70%
RandomForestClassifier	11156	3374	1488	9363	88.23%
KNeighborsClassifier	4562	9968	676	10175	87.09%
GaussianNB	13330	1200	2075	8776	86.53%
SVC	13449	1081	2339	8512	85.18%
RadiusNeighborsClassifier	13318	1212	2367	8484	84.90%
SGDClassifier	3599	10931	674	10177	84.22%
BernoulliNB	12995	1535	2507	8344	83.82%
GradientBoostingClassifier	12645	1885	9605	1246	56.83%



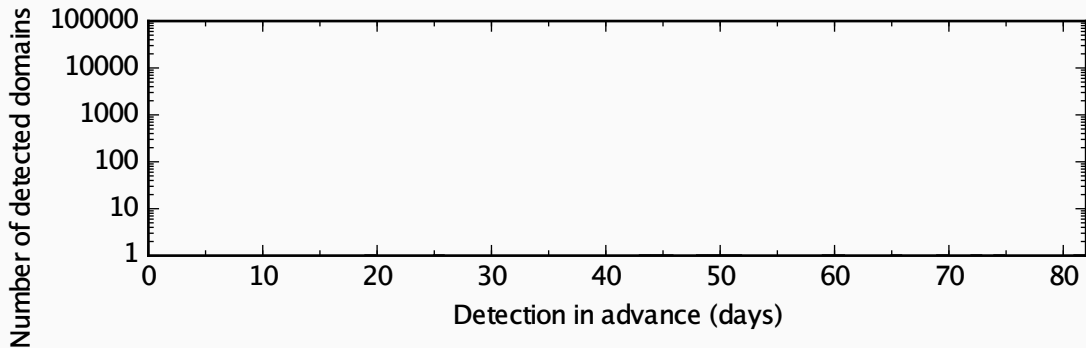
Type	Spam		Ham		Precision
	TP	FN	FP	TN	
AdaBoostClassifier Improved	6688	7842	110	10741	98.38%
AdaBoostClassifier	5971	8559	164	10687	97.32%
MLPClassifier	7273	7257	707	10144	91.14%
DecisionTreeClassifier	6279	8251	695	10156	90.03%
MultinomialNB	12179	2351	1397	9454	89.70%
RandomForestClassifier	11156	3374	1488	9363	88.23%
KNeighborsClassifier	4562	9968	676	10175	87.09%
GaussianNB	13330	1200	2075	8776	86.53%
SVC	13449	1081	2339	8512	85.18%
RadiusNeighborsClassifier	13318	1212	2367	8484	84.90%
SGDClassifier	3599	10931	674	10177	84.22%
BernoulliNB	12995	1535	2507	8344	83.82%
GradientBoostingClassifier	12645	1885	9605	1246	56.83%

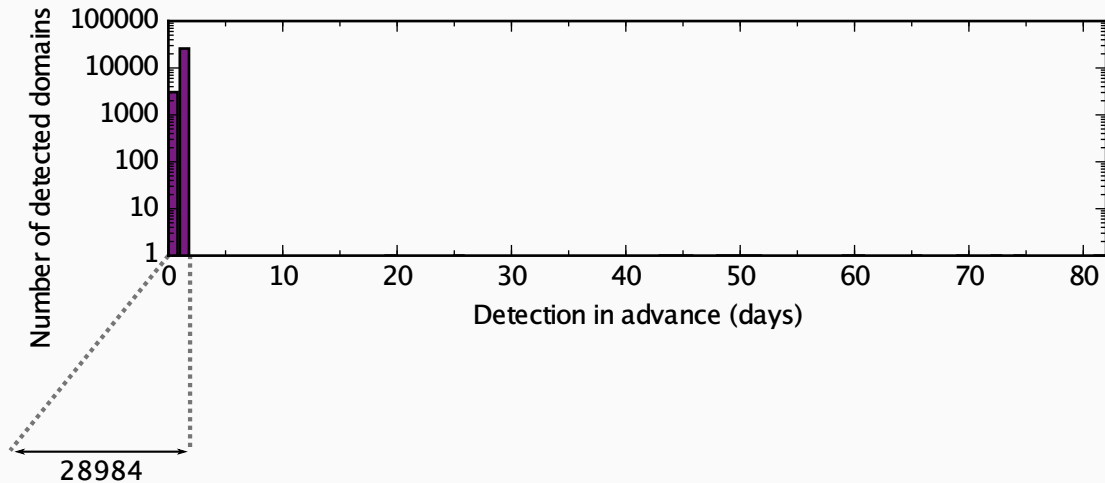


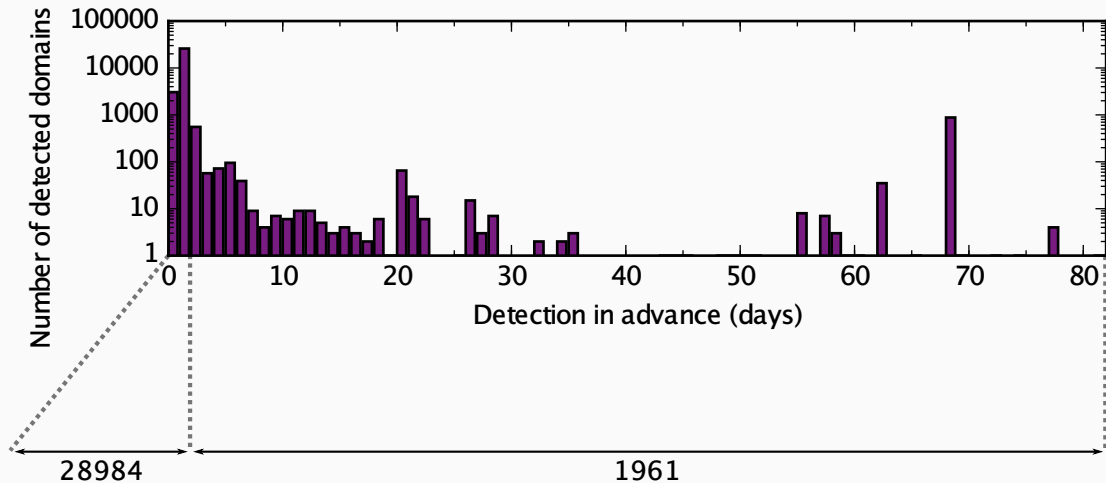
Type	Spam		Ham		Precision
	TP	FN	FP	TN	
AdaBoostClassifier Improved	6688	7842	110	10741	98.38%
AdaBoostClassifier	5971	8559	164	10687	97.32%
MLPClassifier	7273	7257	707	10144	91.14%
DecisionTreeClassifier	6279	8251	695	10156	90.03%
MultinomialNB	12179	2351	1397	9454	89.70%
RandomForestClassifier	11156	3374	1488	9363	88.23%
KNeighborsClassifier	4562	9968	676	10175	87.09%
GaussianNB	13330	1200	2075	8776	86.53%
SVC	13449	1081	2339	8512	85.18%
RadiusNeighborsClassifier	13318	1212	2367	8484	84.90%
SGDClassifier	3599	10931	674	10177	84.22%
BernoulliNB	12995	1535	2507	8344	83.82%
GradientBoostingClassifier	12645	1885	9605	1246	56.83%

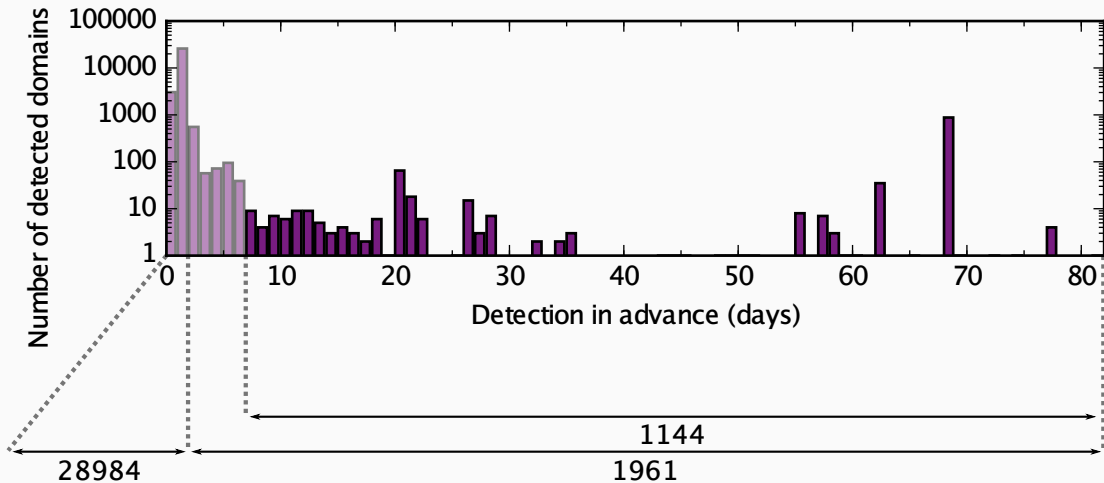
Realtime Blackhole List (RBL)

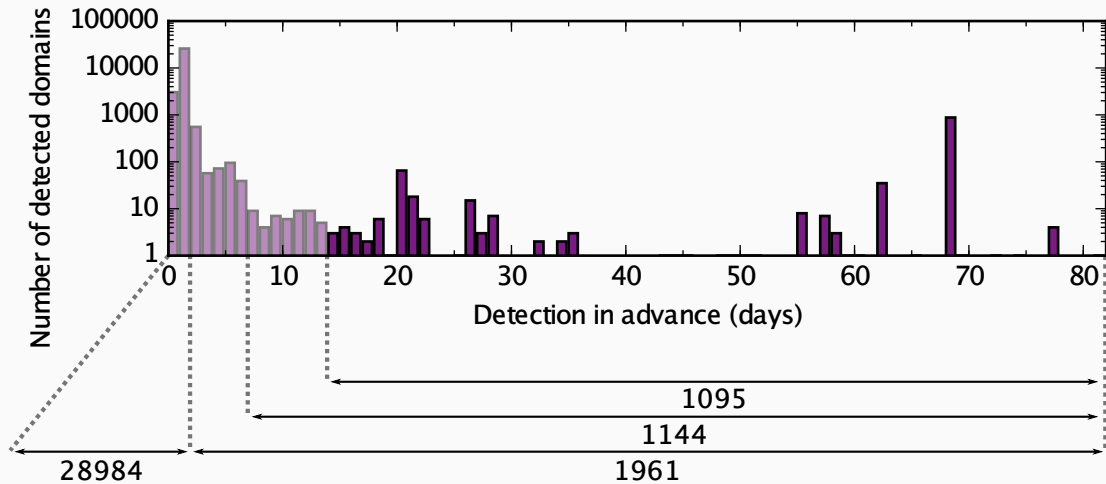


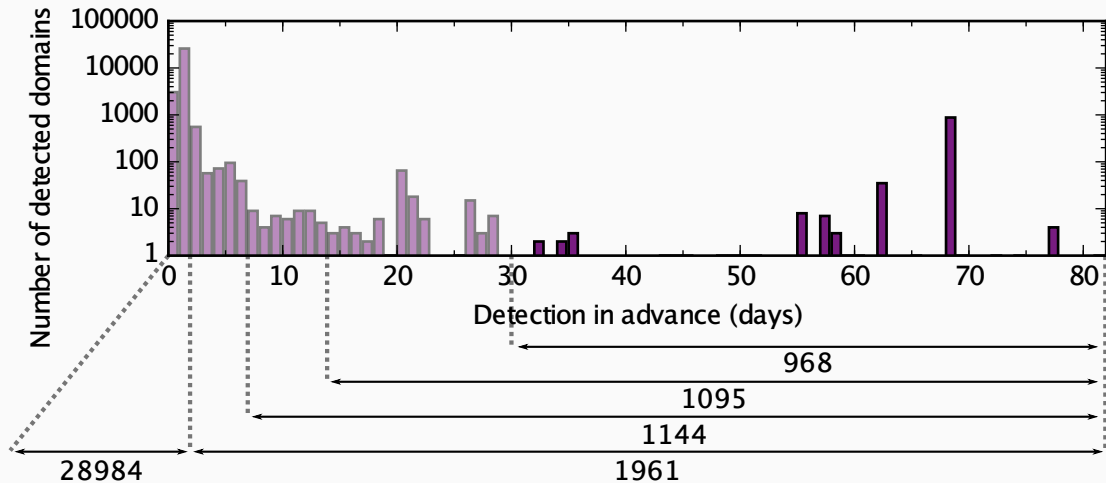


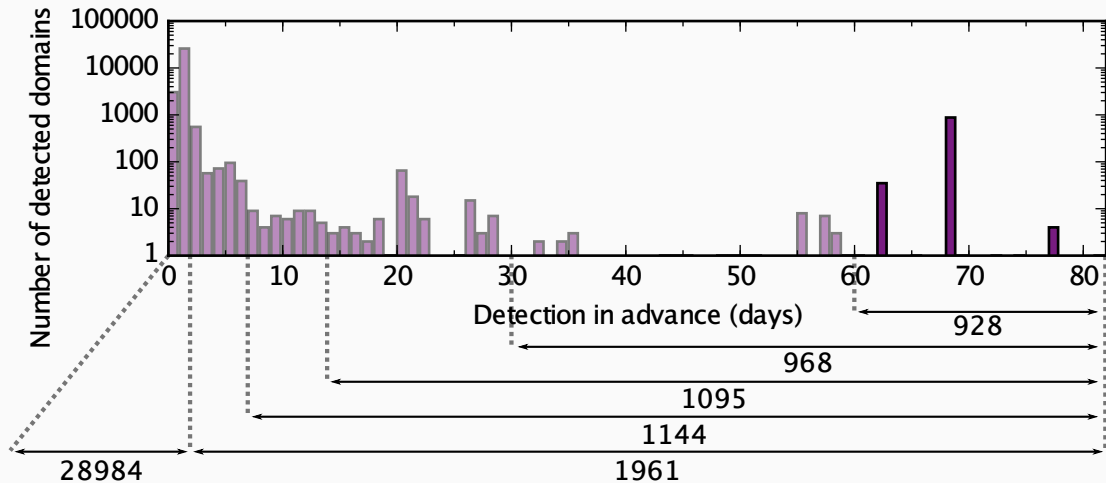


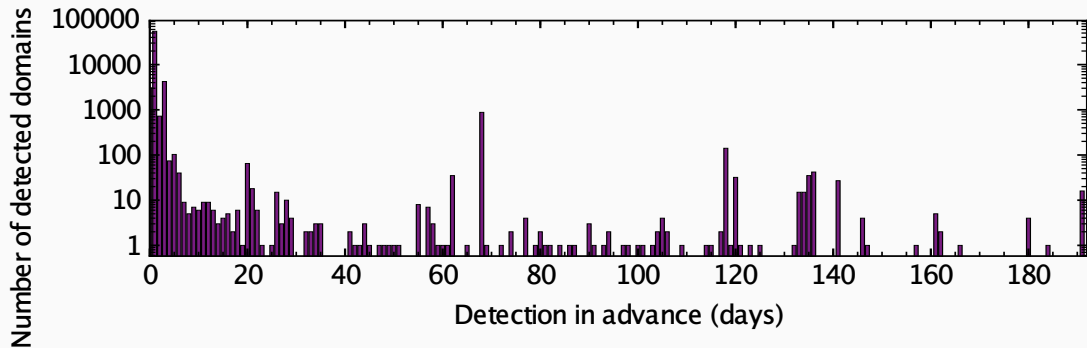


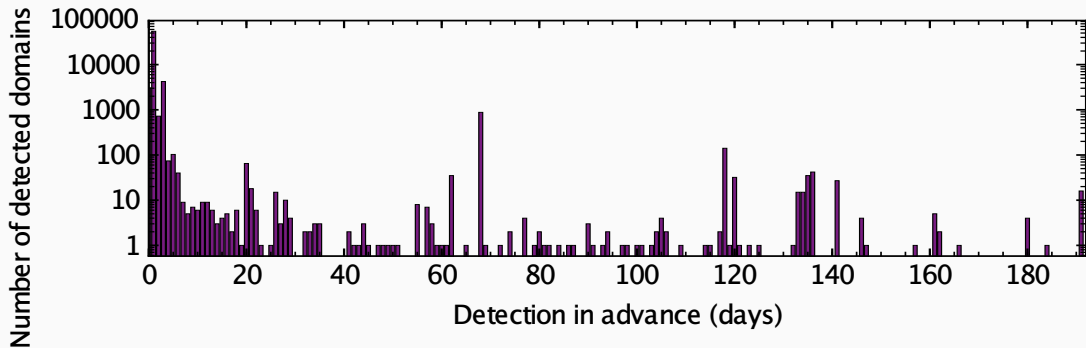




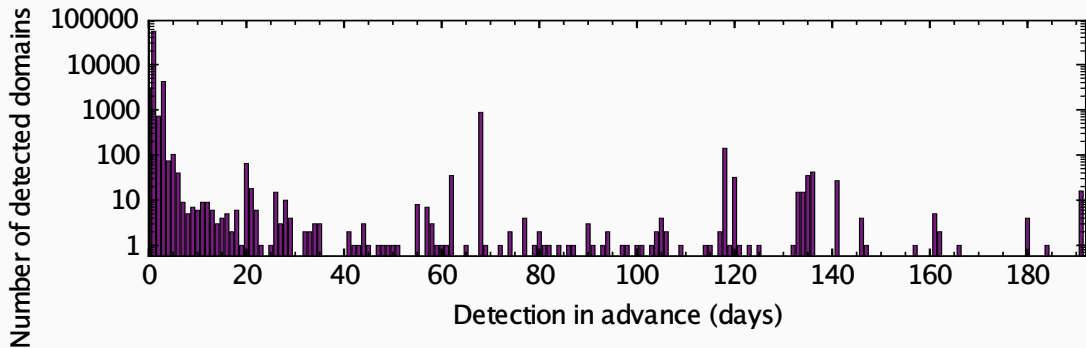




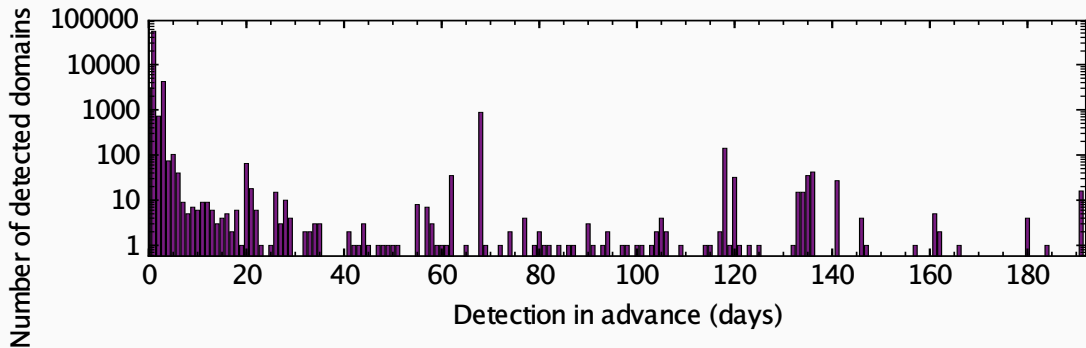




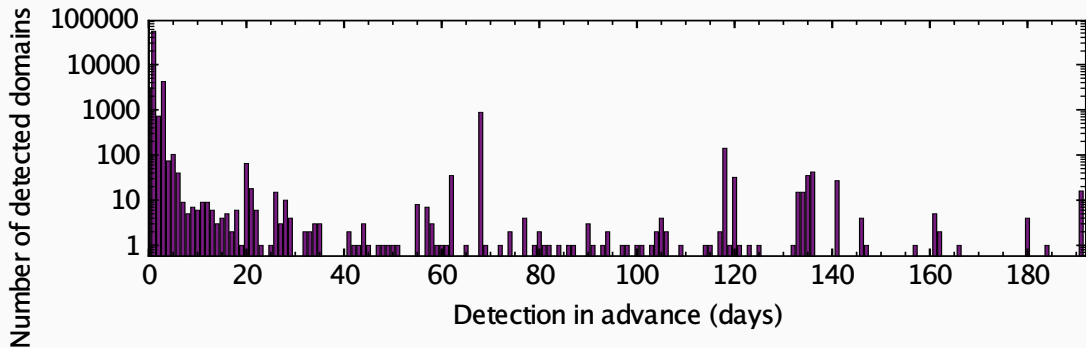
- 112137 domains in total



- 112137 domains in total
- 57724 domains have been detected one day (or less) in advance

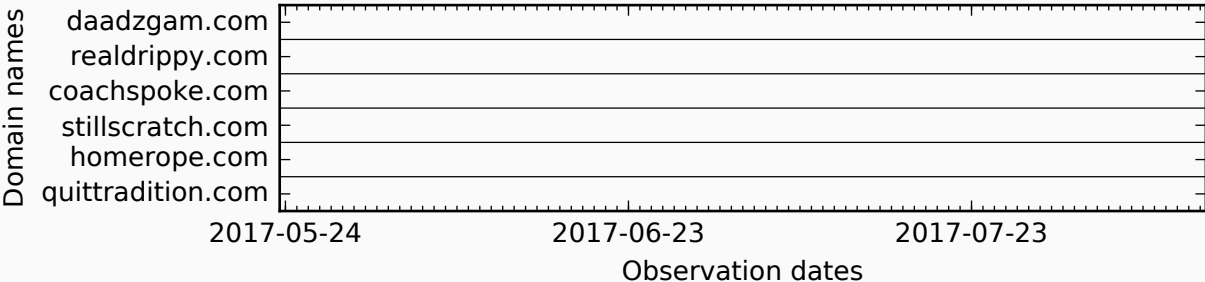


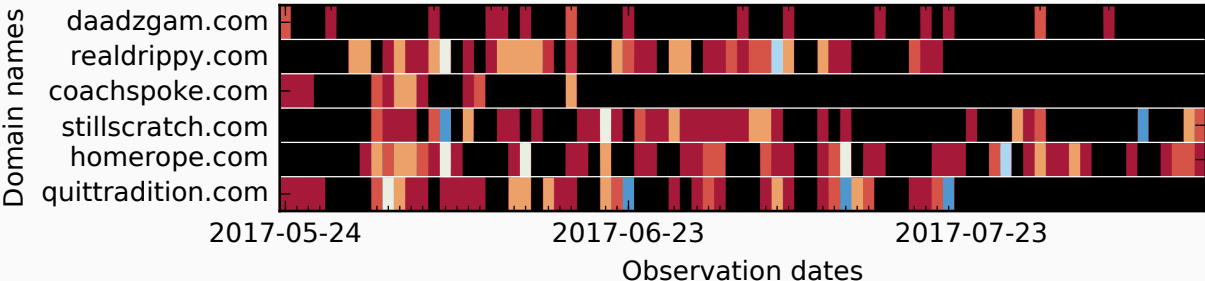
- 112137 domains in total
- 57724 domains have been detected one day (or less) in advance
- 6710 domains have been detected two days (or more) in advance

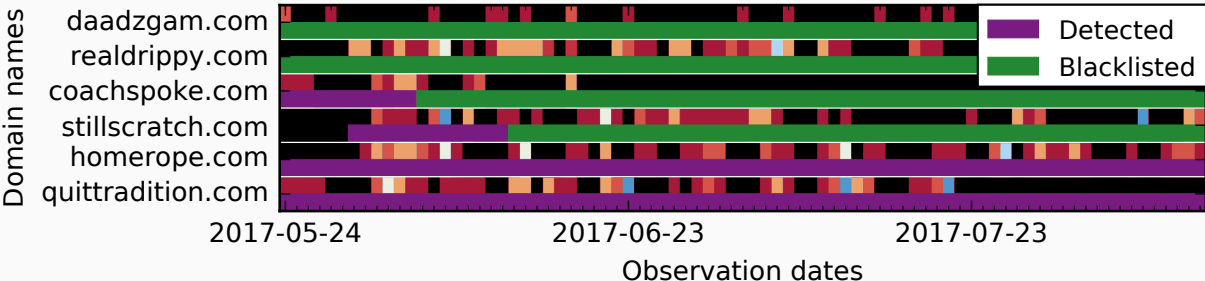


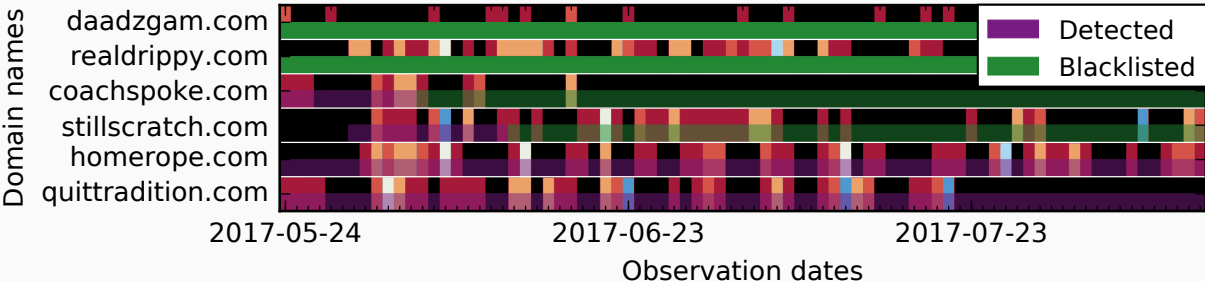
- 112137 domains in total
- 57724 domains have been detected one day (or less) in advance
- 6710 domains have been detected two days (or more) in advance
- 47703 domains have been detected and never blacklisted

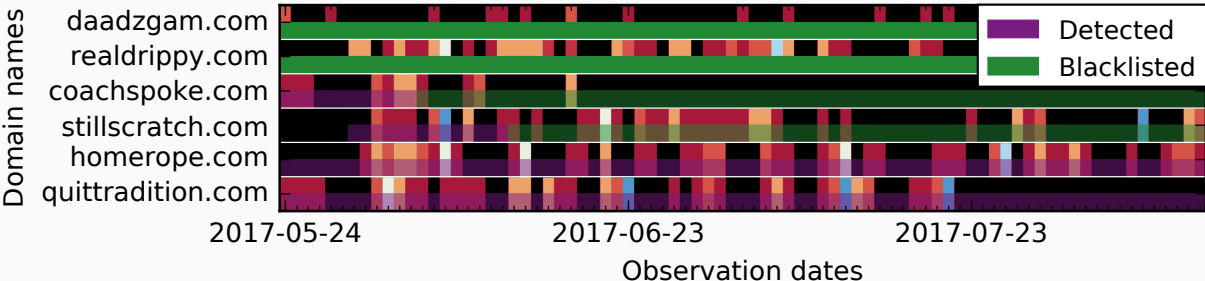




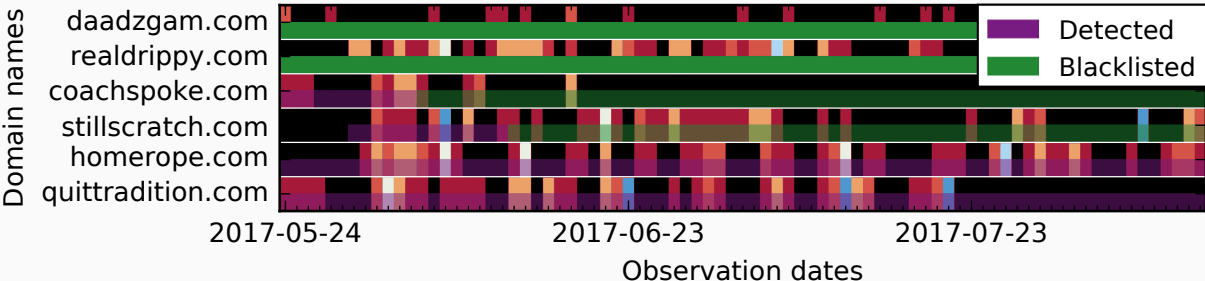




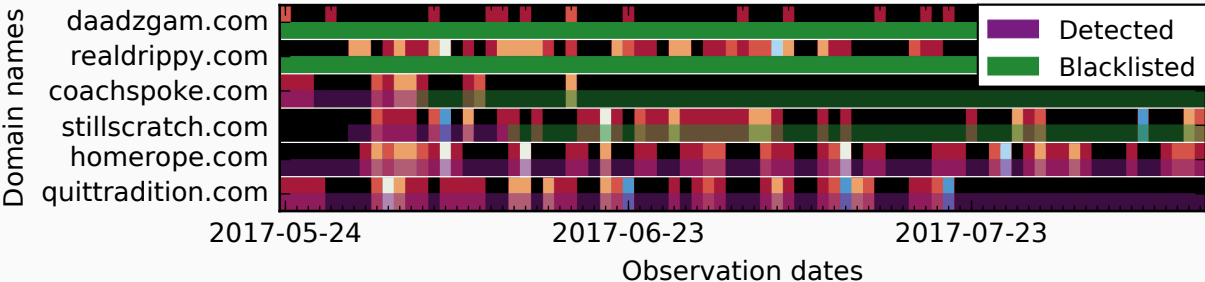




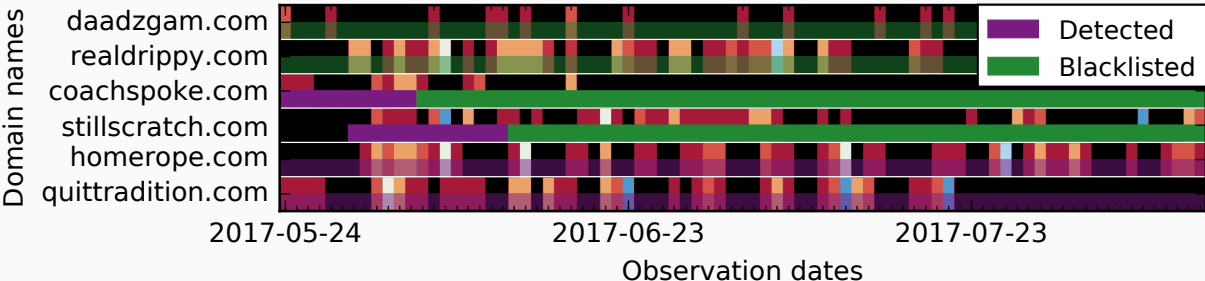
- 20 domains

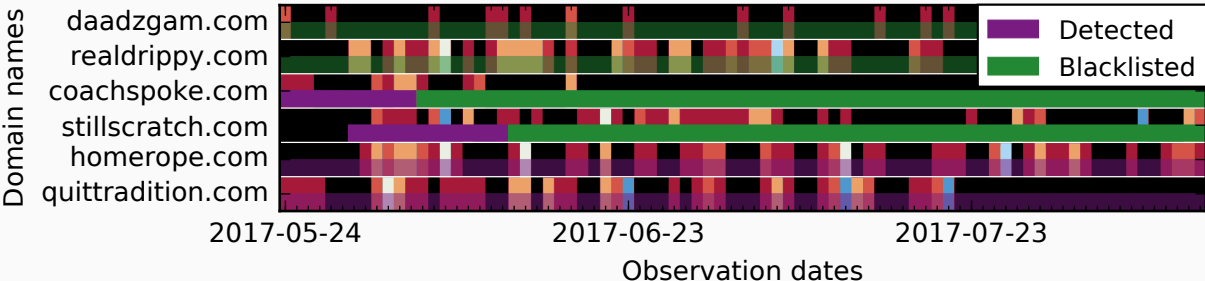


- 20 domains
- 14 (70.00%) domains have an average score of five or higher

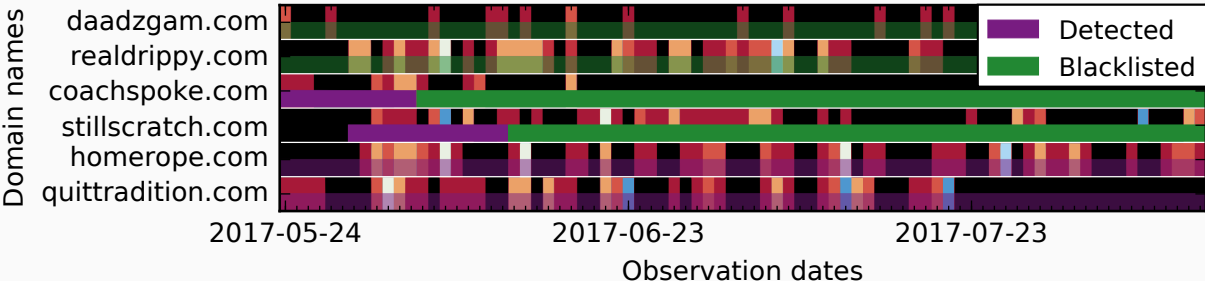


- 20 domains
- 14 (70.00%) domains have an average score of five or higher
- 1188 emails

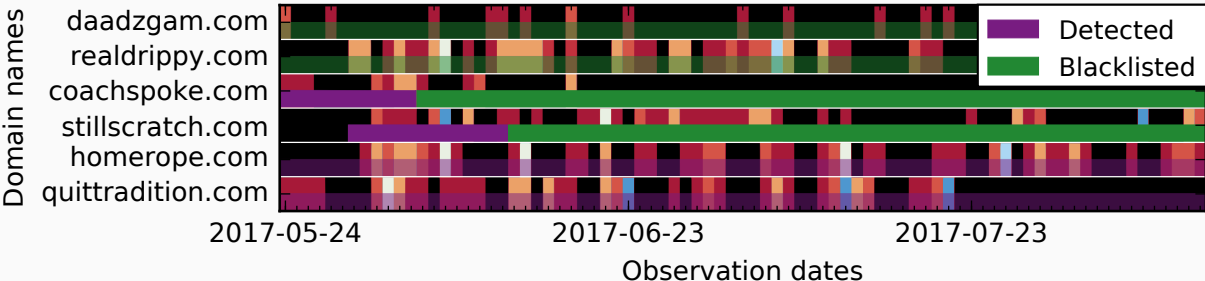




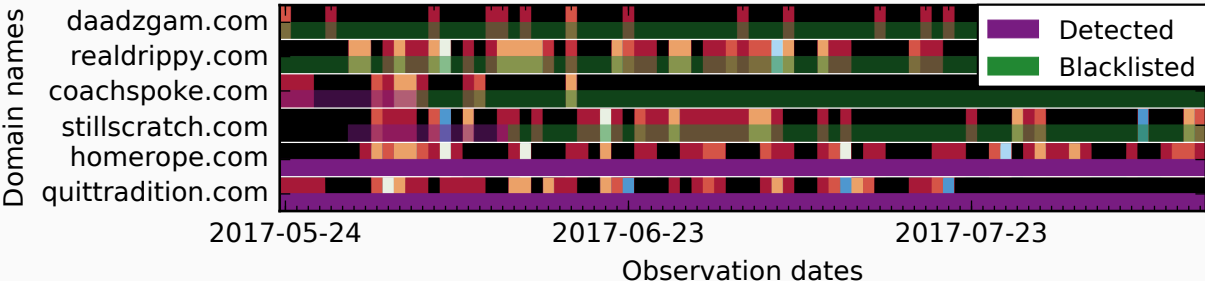
- 29 domains

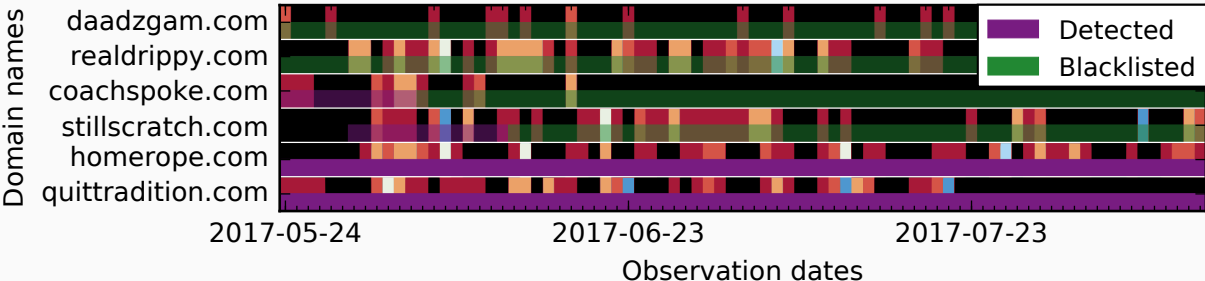


- 29 domains
- 21 (72.41%) domains have an average score of five or higher

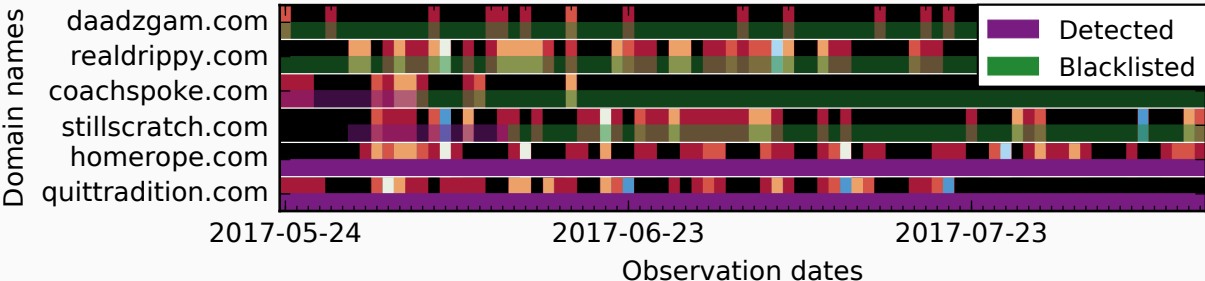


- 29 domains
- 21 (72.41%) domains have an average score of five or higher
- 448 emails

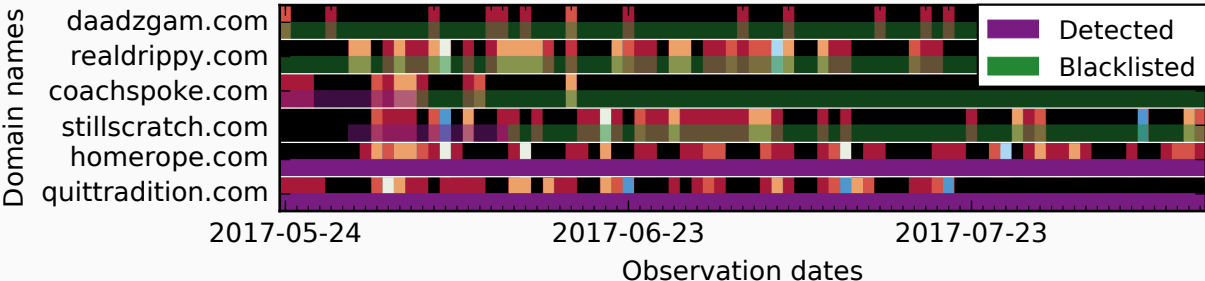




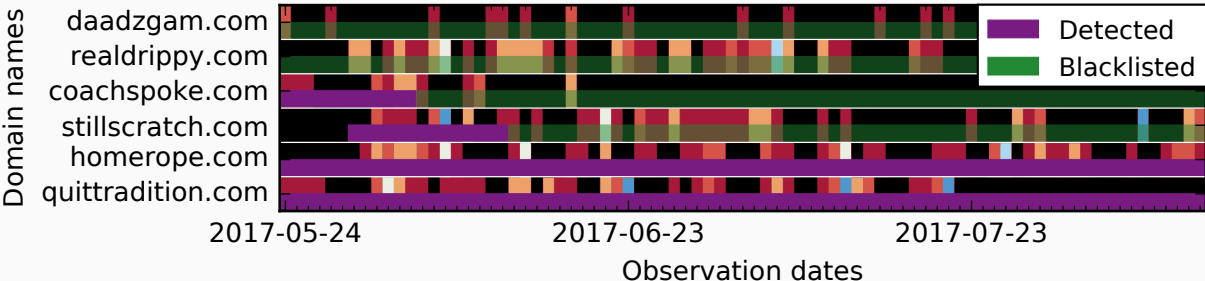
- 64 domains

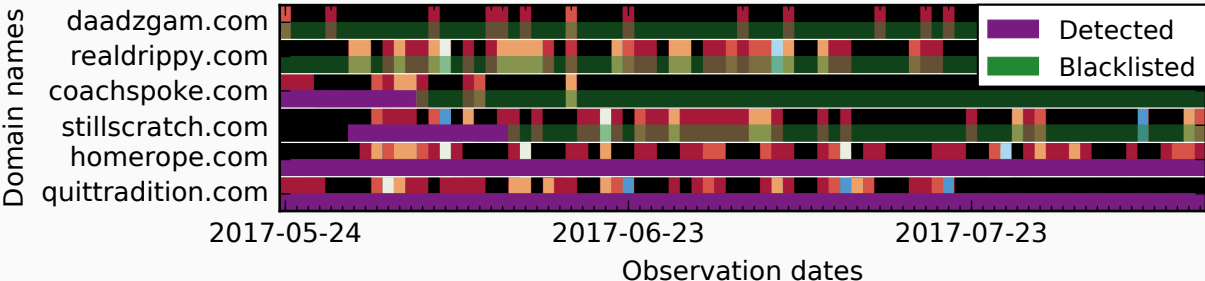


- 64 domains
- 39 (60.93%) domains have an average score of five or higher

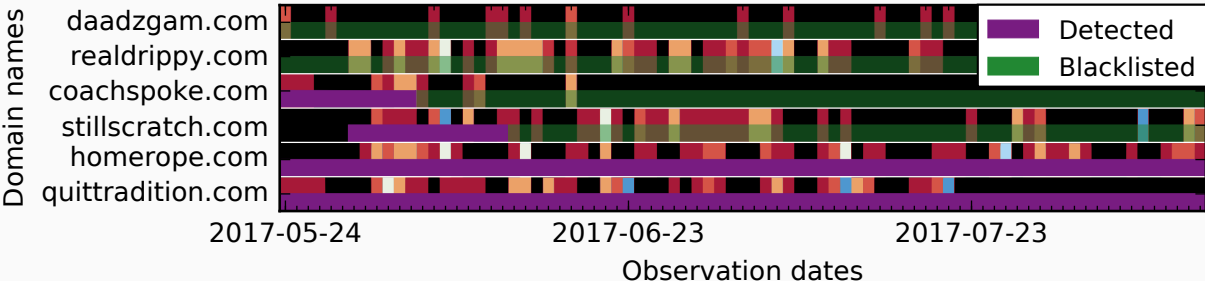


- 64 domains
- 39 (60.93%) domains have an average score of five or higher
- 1006 emails

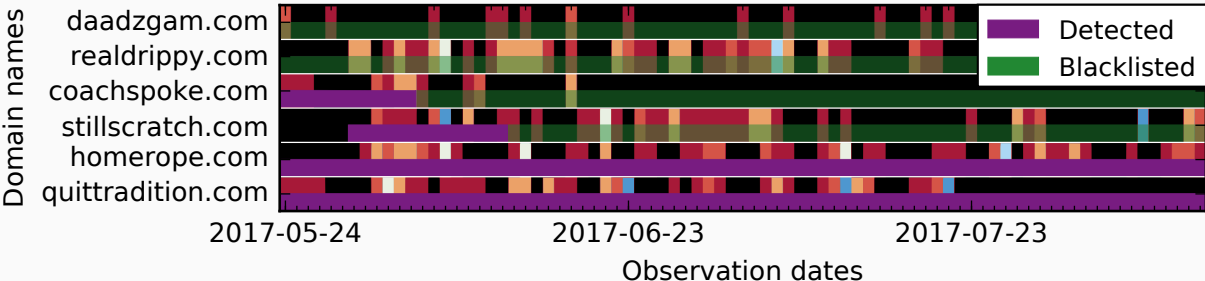




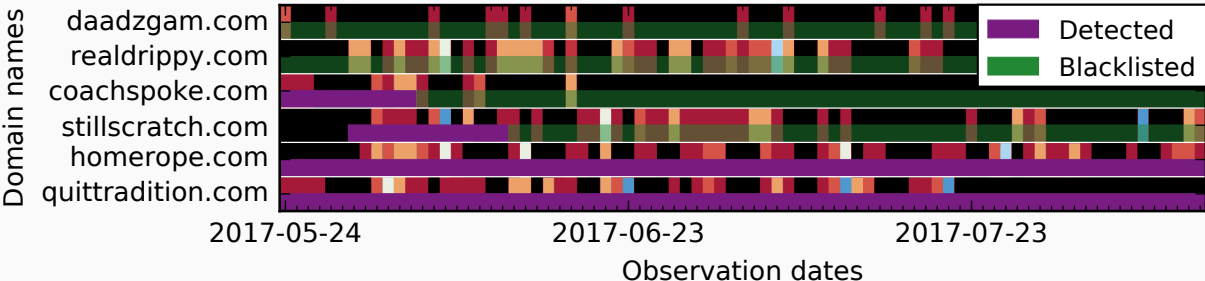
- 1080 emails



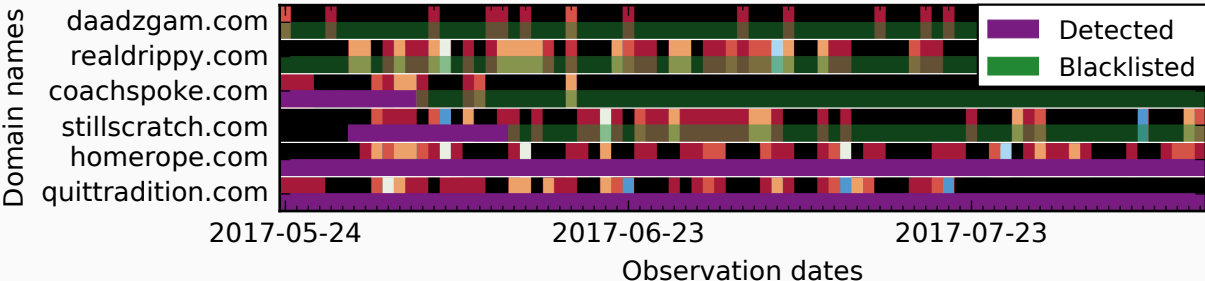
- 1080 emails
- 447 (41.39%) have a score of five or higher



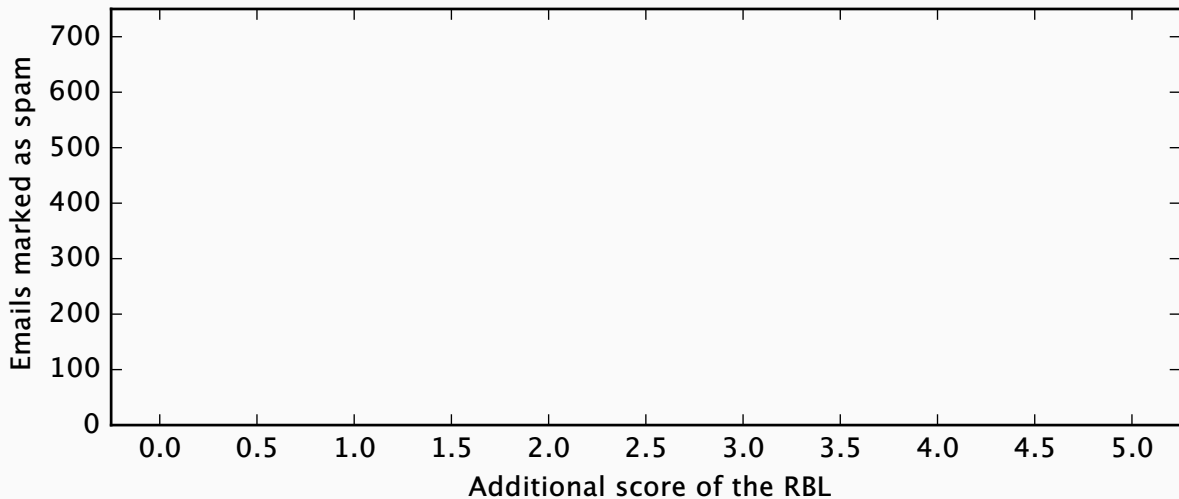
- 633 have a score below five

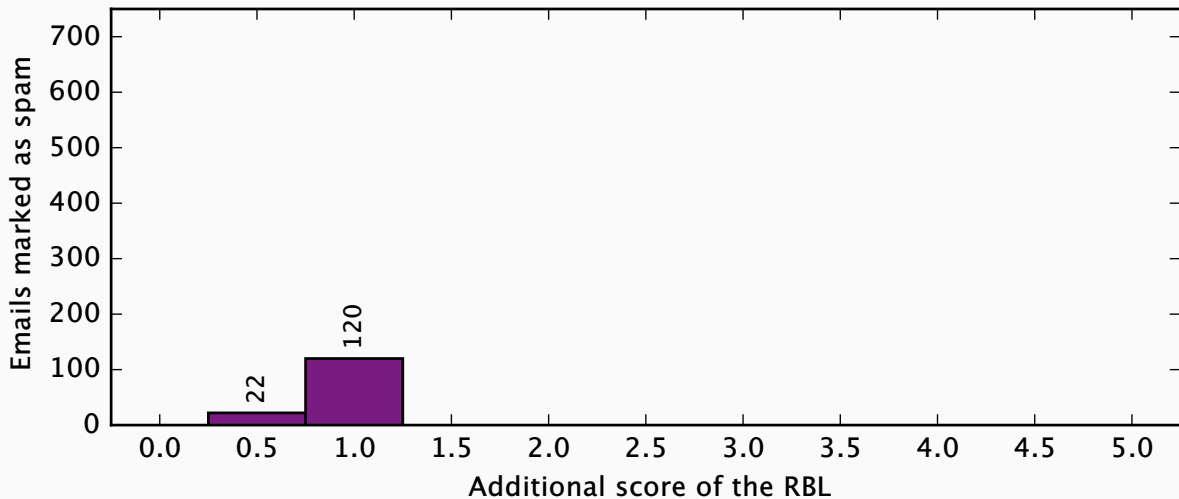


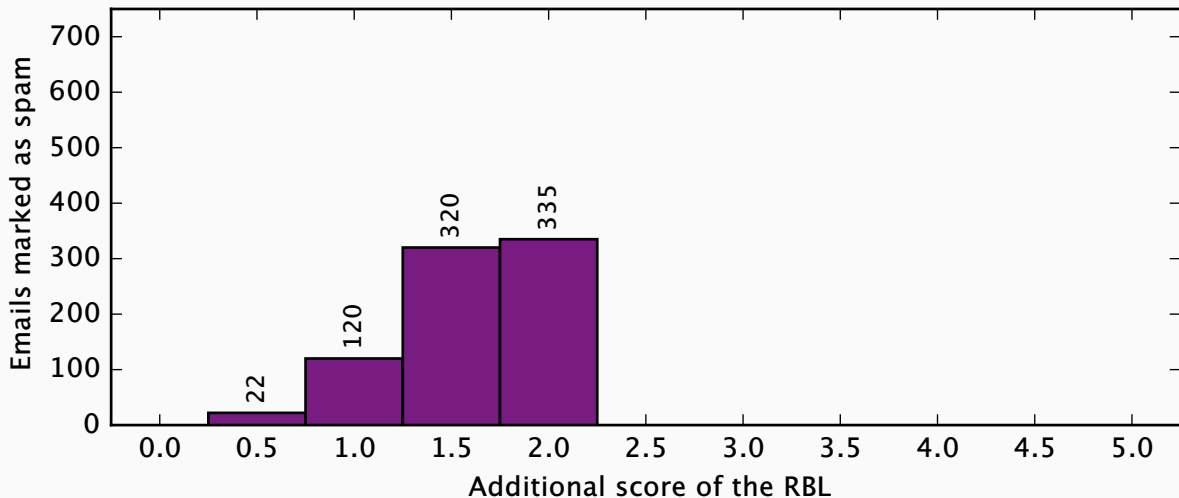
- 633 have a score below five
- counting 52 unique domains

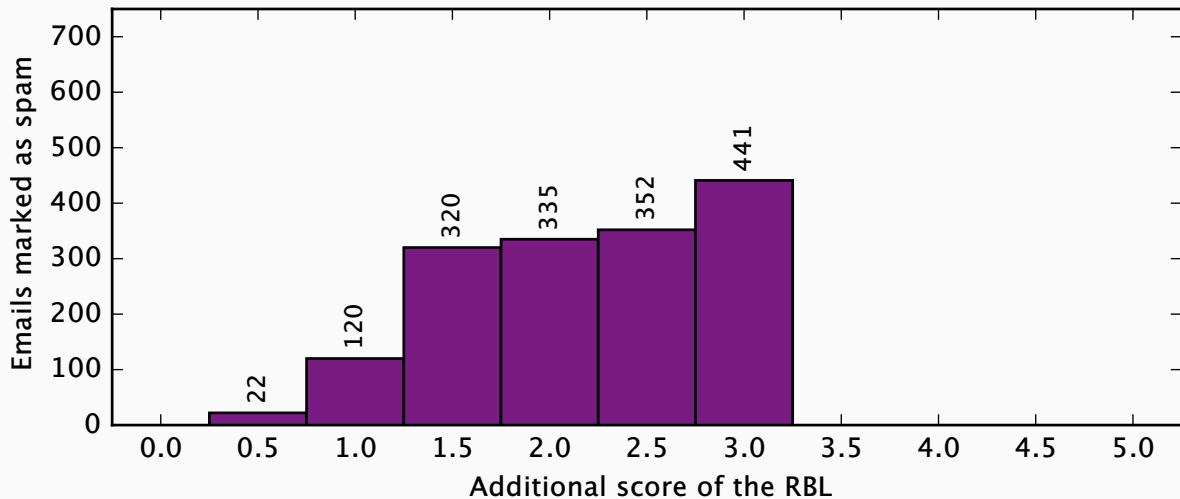


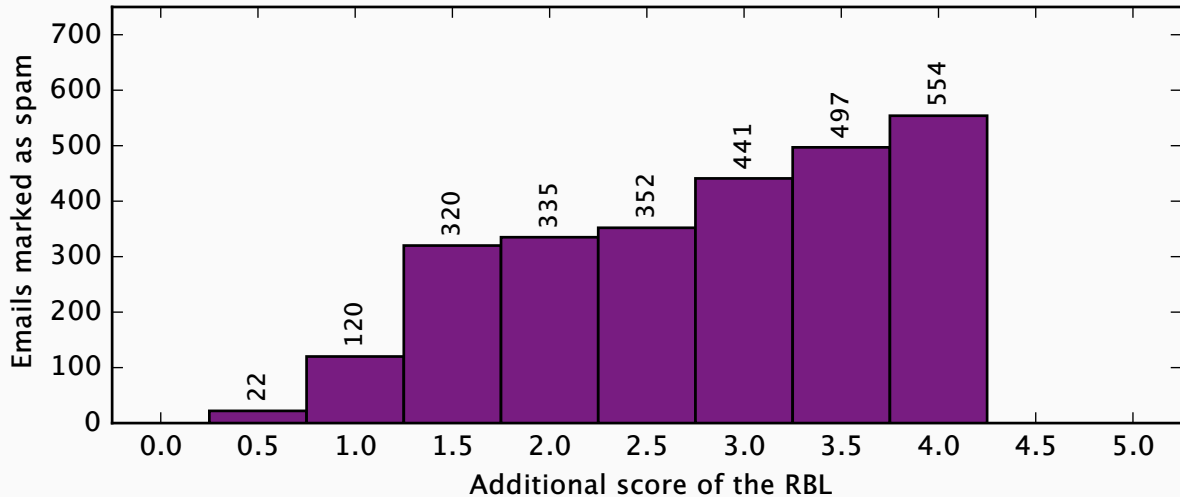
- 633 have a score below five
- counting 52 unique domains
- 13 of these never appear in an email scoring above five

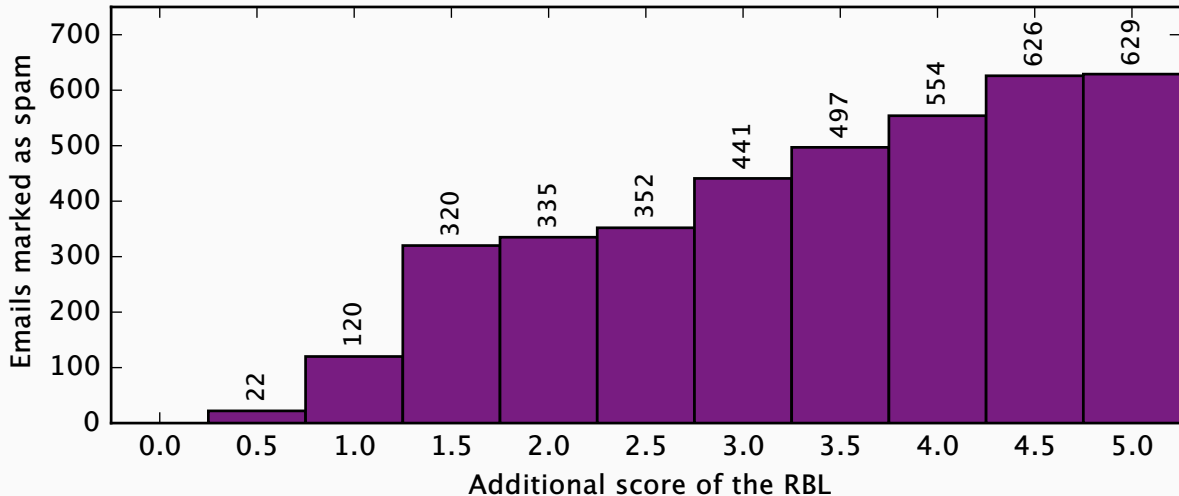


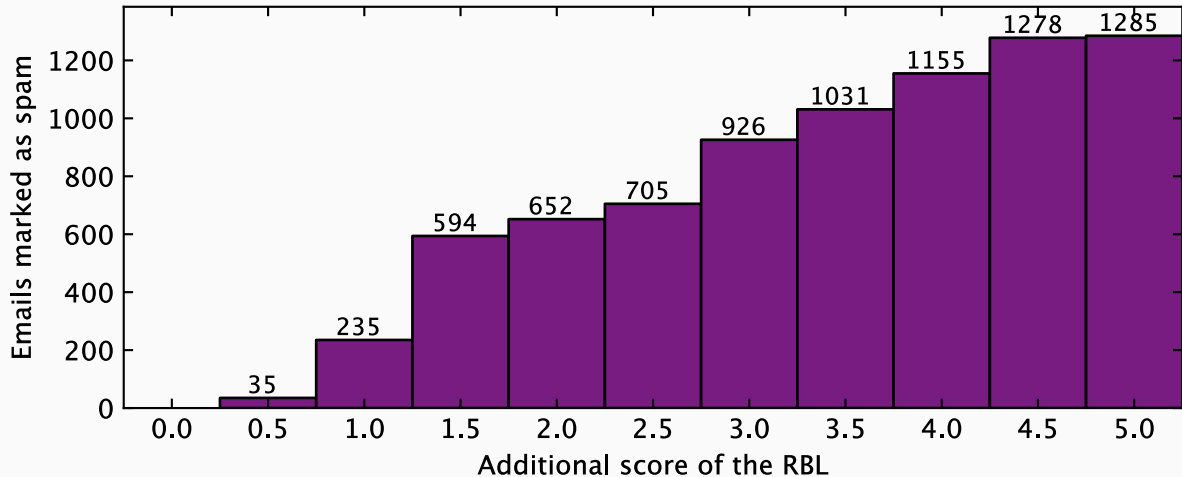














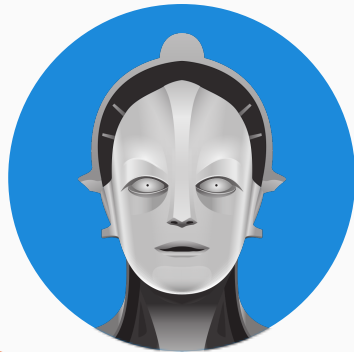
Conclusions



What is the advantage of proactive
snowshoe spam domain detection
using DNS data?



Conclusions



Conclusions



o.i.vandertoorn@utwente.nl



