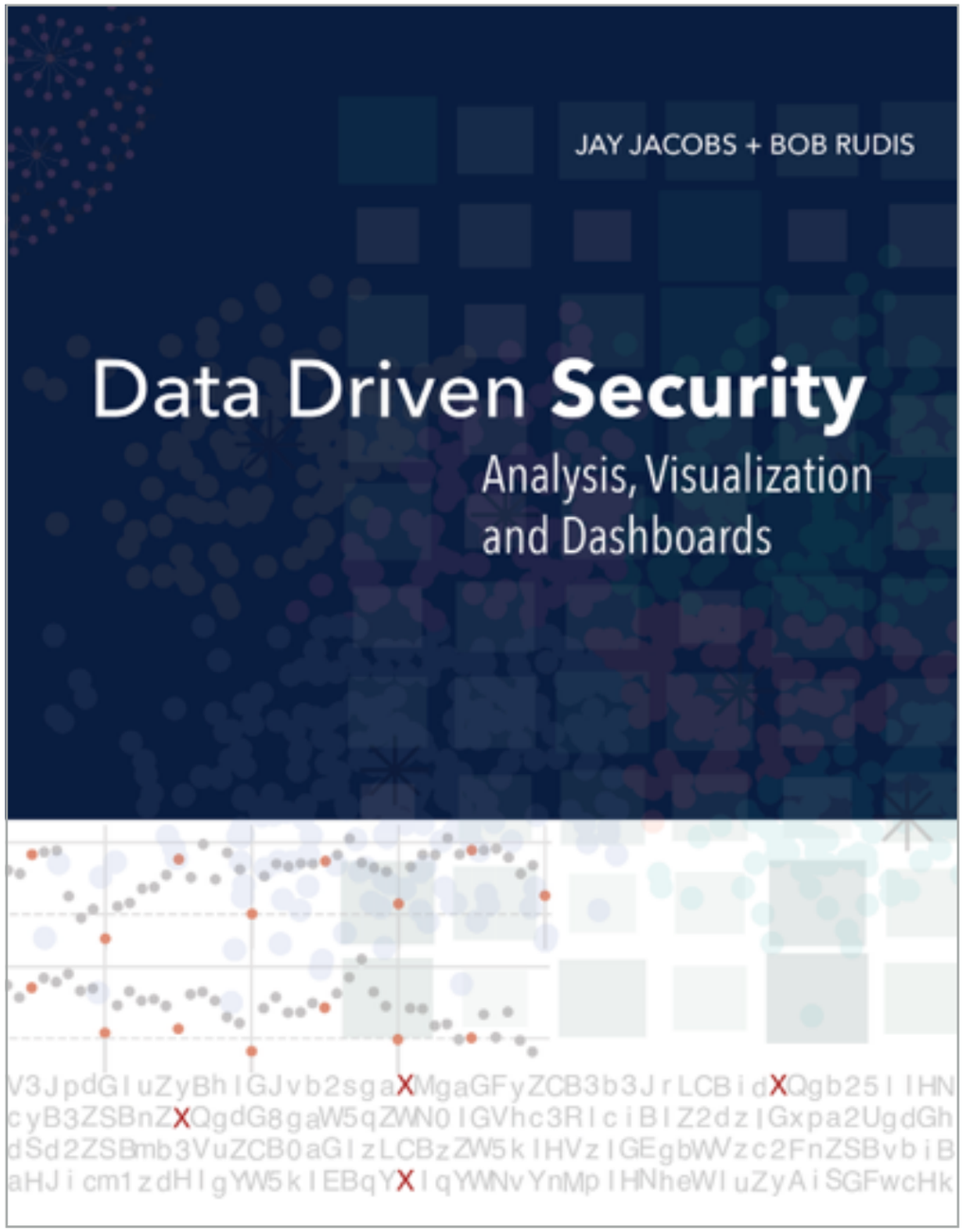
Detecting **Algorithmically Generated Domains**

Jay Jacobs
March 7th, 2015
@jayjacobs

Jay Jacobs

Data Driven **Security**

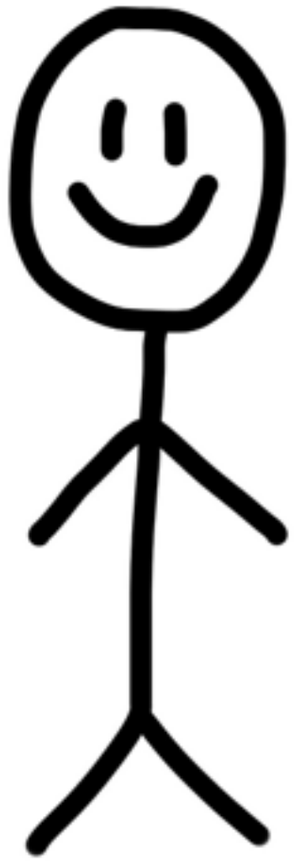


http://www.verizonenterprise.com/DBIR/

http://amzn.to/ddsec

IP and Domains

Data Driven Security



Where is wikipedia.com?

208.80.154.224

DNS

Hello 208.80.154.224

208.80.154.224

IP and Domains

Data Driven Security

The early days…

Malicious Software

201.130.44.18

216.58.216.238

173.194.46.86

badserver1.com

badserver2.com

badserver3.com

Domain Generating Algorithms (DGA)

Algorithms that generate pseudo-random domain names. Used by malware to communicate with a controlling server.

IP and Domains

Data Driven Security

And Now…

Malicious Software

xautsghmdtoofnn.com	No
exotugfsphafhxt.com	No
kykwdvibpsccedj.com	No
nmexqpgdqnxysvp.com	No
civtuqeeoqueg.com	No
kmlnfcegedkle.com	No
uailugreagmxe.com	No
cohbwhwwdrqqv.com	No
dbbsxhljgdexb.com	No
vyqabtoayoheu.com	Yes

Building a Classifier

Data Driven Security

1. Obtain labelled data
2. Add or derive features
3. Create test/train data sets
4. Run classification algorithm[s] on training data
5. Evaluate model and features on test data

The Challenge

Data Driven Security

Given a list of domains, classify them as either benign or DGA

hrjiojexjpjodkk.ru	1pbwqpi1599j8ppvlrrl1u580qx.com
xautsghmdtoofnn.ru	1k8le0smhotv93pt7c295rr6e.net
exotugfsphafhxt.ru	1yzkujin8r9xi1mt6ku0bt5nll.biz
kykwdvibpsccedj.ru	xkvy3k2afwwm1baincm1qds8sz.org
nmexqpgdqnxysvp.ru	ophh561k6a4vl3dotmo9fshq5.com
civtuqeeoqueg.ru	1jf7fx5113uv7obiu9k0mqfgyn.net
kmlnfcegedkle.ru	1e9aiud11ed98c1x4gnp7hi4q8s.org
uailugreagmxe.ru	1drt4la5ndhno93q70cmnwmph.net
cohbwhwwdrqqv.ru	6e6bqs1bbmcovukufk3hzhri0.com
dbbsxhljgdexb.ru	1nh8no0q31xv313hzv6xmskrm6.org

3k-5k (3.7k median) malicious domains registered daily in US domains,
about 1% of US domains are malicious

The Solution

Data Driven Security

For Python:

https://github.com/ClickSecurity/data_hacking [dga_detection]

For R:

https://github.com/jayjacobs/dga-tutorial

https://github.com/jayjacobs/dga

http://datadrivensecurity.info/blog/posts/2014/Sep/dga-part1/

Both locations have labelled data!

Building a Classifier

Data Driven Security

1. ~~Obtain labelled data~~
2. **Add or derive features**
3. Create test/train data sets
4. Run classification algorithm[s] on training data
5. Evaluate model and features on test data

Features?

Data Driven **Security**

```
facebook.com
doubleclick.com
google-analytics.com
akamaihd.com
twitter.com
youtube.com
scorecardresearch.com
microsoft.com
googletagservices.com
msftncsi.com
```

```
1pbwqpi1599j8ppvlrrl1u580qx.com
1k8le0smhotv93pt7c295rr6e.net
1yzkujin8r9xi1mt6ku0bt5nll.biz
xkvy3k2afwwm1baincm1qds8sz.org
ophh561k6a4vl3dotmo9fshq5.com
1jf7fx5113uv7obiu9k0mqfgyn.net
1e9aiud11ed98c1x4gnp7hi4q8s.org
1drt4la5ndhno93q70cmnwmph.net
6e6bqs1bbmcovukufk3hzhri0.com
1nh8no0q31xv313hzv6xmskrm6.org
```

Features

- Length
- Entropy
- n-grams
- dictionary word matches
- and Others

```
facebook.com
doubleclick.com
google-analytics.com
1pbwqpi1599j8ppvlrrl1u580qx.com
1k8le0smhotv93pt7c295rr6e.net
1yzkujin8r9xi1mt6ku0bt5nll.biz
```

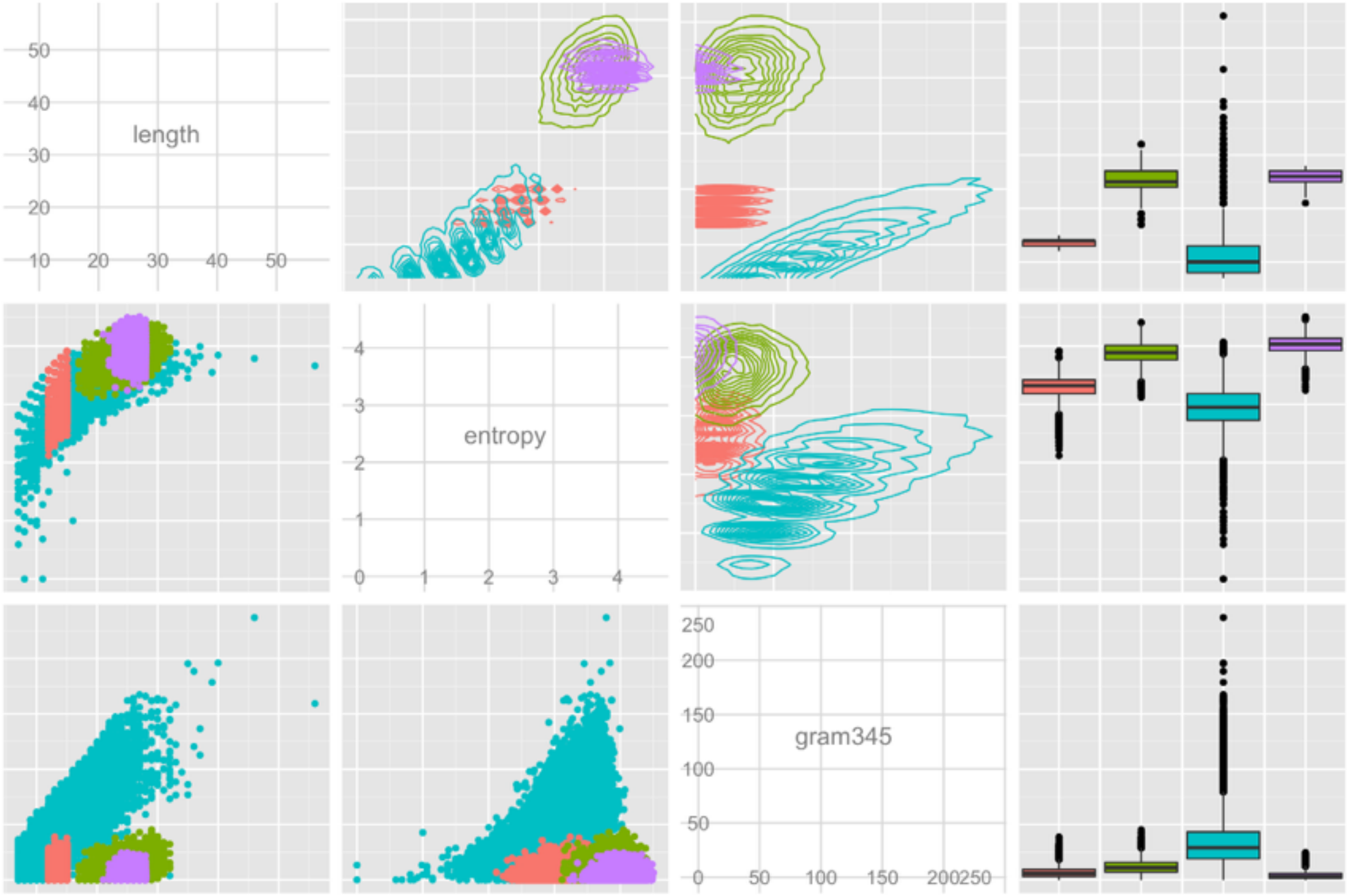
n-grams

Data Driven Security

3-grams		4-grams		4-grams (dga)	
ing	3.5482	line	3.2148	eqkb	1.0791
lin	3.4924	nlin	3.1595	inqo	1.0791
ine	3.4743	onli	3.1513	wgyh	1.0791
ter	3.3857	tion	3.1322	gylr	1.0413
ion	3.3727	free	3.0330	inqw	1.0413
ent	3.3471	news	3.0051	ljtg	1.0413
tor	3.3469	mark	2.9995	lrdq	1.0413
the	3.3031	game	2.9978	nbyp	1.0413
ers	3.2736	dire	2.9845	onba	1.0413
tra	3.2683	irec	2.9781	qcws	1.0413

domain	class	length	entropy	onegram	threegram	fourgram	fivegram	gram345
facebook	legit	8	2.750000	36.93176	15.66067	10.39223	6.844194	32.89709
google-analytics	legit	16	3.500000	74.47313	32.33994	16.50915	11.601353	60.45045
akamaihd	legit	8	2.405639	37.22381	11.01290	1.50515	0.000000	12.51805
facebook	legit	8	2.750000	36.93176	15.66067	10.39223	6.844194	32.89709
microsoft	legit	9	2.947703	42.15909	17.11639	11.39665	7.493930	36.00697
googletagservices	legit	17	3.292770	79.98536	36.45091	23.18288	12.778621	72.41240

domain	class	length	entropy	onegram	threegram	fourgram	fivegram	gram345
exotugfsphafhxt	dga	15	3.373557	67.02298	8.673246	0	0	8.673246
civtuqeeoqueg	dga	13	3.026987	57.67474	8.827826	0	0	8.827826
cohbwhwwdrqqv	dga	13	3.026987	54.43738	0.000000	0	0	0.000000
qixyfrsfiyied	dga	13	3.026987	57.37876	9.761103	0	0	9.761103
ptyjwsefmtslk	dga	13	3.392747	58.05692	4.670913	0	0	4.670913
hvuwoxwkfpbwy	dga	13	3.334679	55.16979	0.000000	0	0	0.000000


length
entropy
gram345
yptolocker
goz
subclass
legit
newgoz

Building a Classifier

Data Driven Security

1. ~~Obtain labelled data~~
2. ~~Add or derive features~~
3. **Create test/train data sets**
4. Run classification algorithm[s] on training data
5. Evaluate model and features on test data

Train Data

Data Driven Security

Python

```
# Hold out 10%
hold_out_alexa = alexa_dataframe[alexa_total*.9:]
alexa_dataframe = alexa_dataframe[:alexa_total*.9]
hold_out_dga = dga_dataframe[dga_total*.9:]
dga_dataframe = dga_dataframe[:dga_total*.9]

# Pull together our hold out set
hold_out_domains = pd.concat([hold_out_alexa, hold_out_dga], ignore_index=True)

# Concatenate the domains in a big pile!
all_domains = pd.concat([alexa_dataframe, dga_dataframe], ignore_index=True)
```

R

```
library(caret)

testindex <- createDataPartition(dgadomain$subclass, p=0.9, list=F)
traindga <- dgadomain[testindex, ]
testdga <- dgadomain[-testindex, ]
```

Building a Classifier

Data Driven Security

1. ~~Obtain labelled data~~
2. ~~Add or derive features~~
3. ~~Create test/train data sets~~
4. **Run classification algorithm[s] on training data**
5. Evaluate model and features on test data

Run Classifer(s)

Data Driven Security

Python

```
# List of feature vectors (scikit learn uses 'X' for the matrix of feature vectors)
X = all_domains.as_matrix(['length', 'entropy'])

# Labels (scikit learn uses 'y' for classification labels)
y = np.array(all_domains['class'].tolist()) # Yes, this is weird but it needs
                                            # to be an np.array of strings
```

```
import sklearn.ensemble
clf = sklearn.ensemble.RandomForestClassifier(n_estimators=20,compute_importances=True)

# Train on a 80/20 split
from sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```

Run Classifer(s)

Data Driven Security

R

```
# set up Train Control
ctrl <- trainControl(method="repeatedcv", repeats=5, classProbs=TRUE,
                     summaryFunction=twoClassSummary)

# Random Forest
rfFit <- train(class ~ ., data = traindga, metric="ROC",
               method = "rf", trControl = ctrl)
```

```
# Support Vector Machine
svmFit <- train(class ~ ., data = traindga, method = "svmRadial",
                preProc = c("center", "scale"), metric="ROC",
                tuneLength = 10, trControl = ctrl)

# C4.5
c45Fit <- train(class ~ ., data = traindga, method = "J48",
                metric="ROC", trControl = ctrl)
```

Building a Classifier

Data Driven Security

1. ~~Obtain labelled data~~
2. ~~Add or derive features~~
3. ~~Create test/train data sets~~
4. ~~Run classification algorithm[s] on training data~~
5. **Evaluate model and features on test data**

Evaluate Model

Data Driven Security

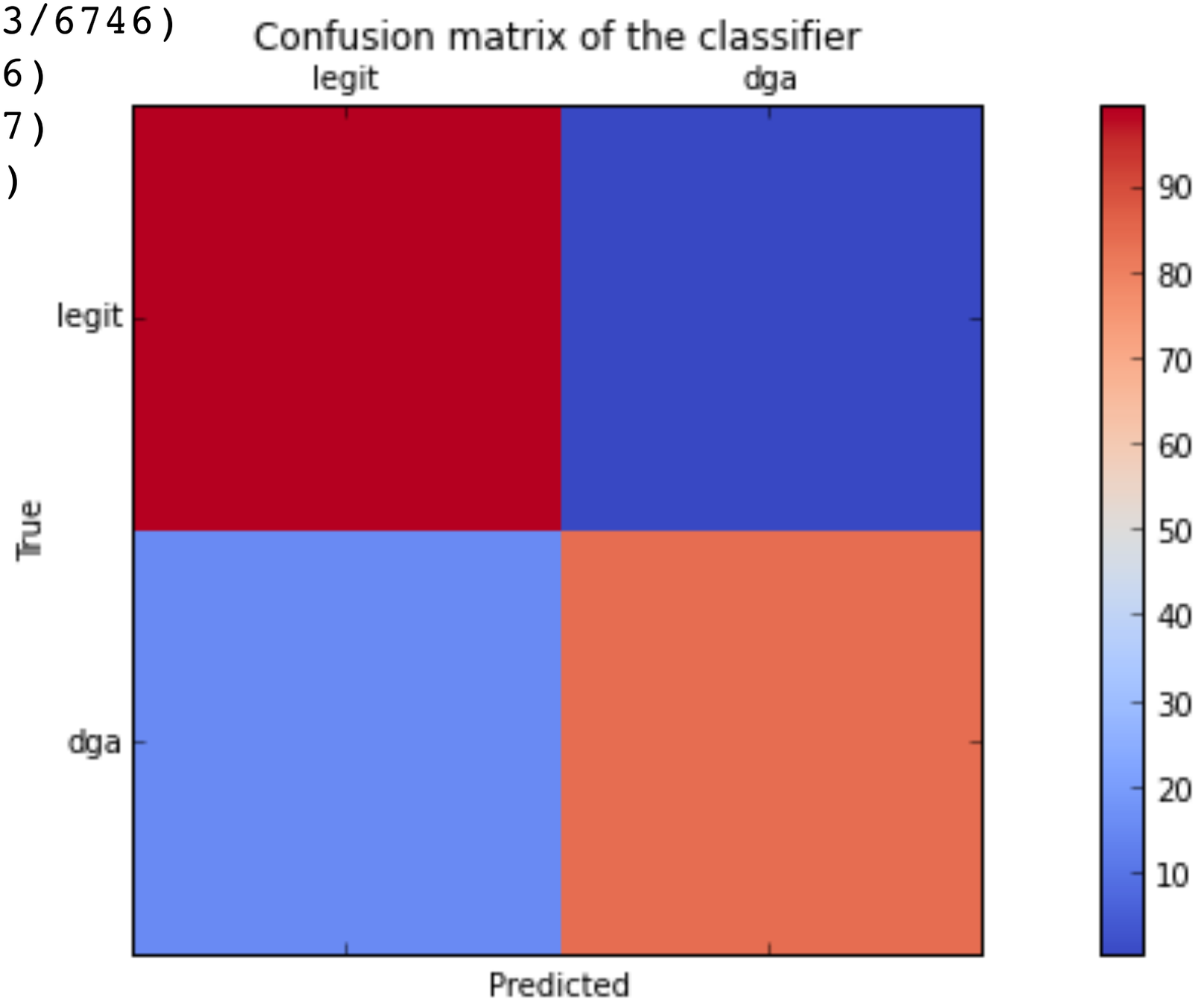
R

```
          Reference
Prediction   dga legit
     dga   39292   282
     legit   206 64458

               Accuracy : 0.9953
                 95% CI : (0.9949, 0.9957)
    No Information Rate : 0.6211
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9869
 Mcnemar's Test P-Value : 0.0006861

            Sensitivity : 0.9948
            Specificity : 0.9956
         Pos Pred Value : 0.9929
         Neg Pred Value : 0.9968
             Prevalence : 0.3789
         Detection Rate : 0.3769
   Detection Prevalence : 0.3797
      Balanced Accuracy : 0.9952
```

Evaluate Model

Data Driven **Security**

```
Confusion Matrix Stats
legit/legit: 99.51% (6713/6746)
legit/dga: 0.49% (33/6746)
dga/legit: 15.73% (42/267)
dga/dga: 84.27% (225/267)
```

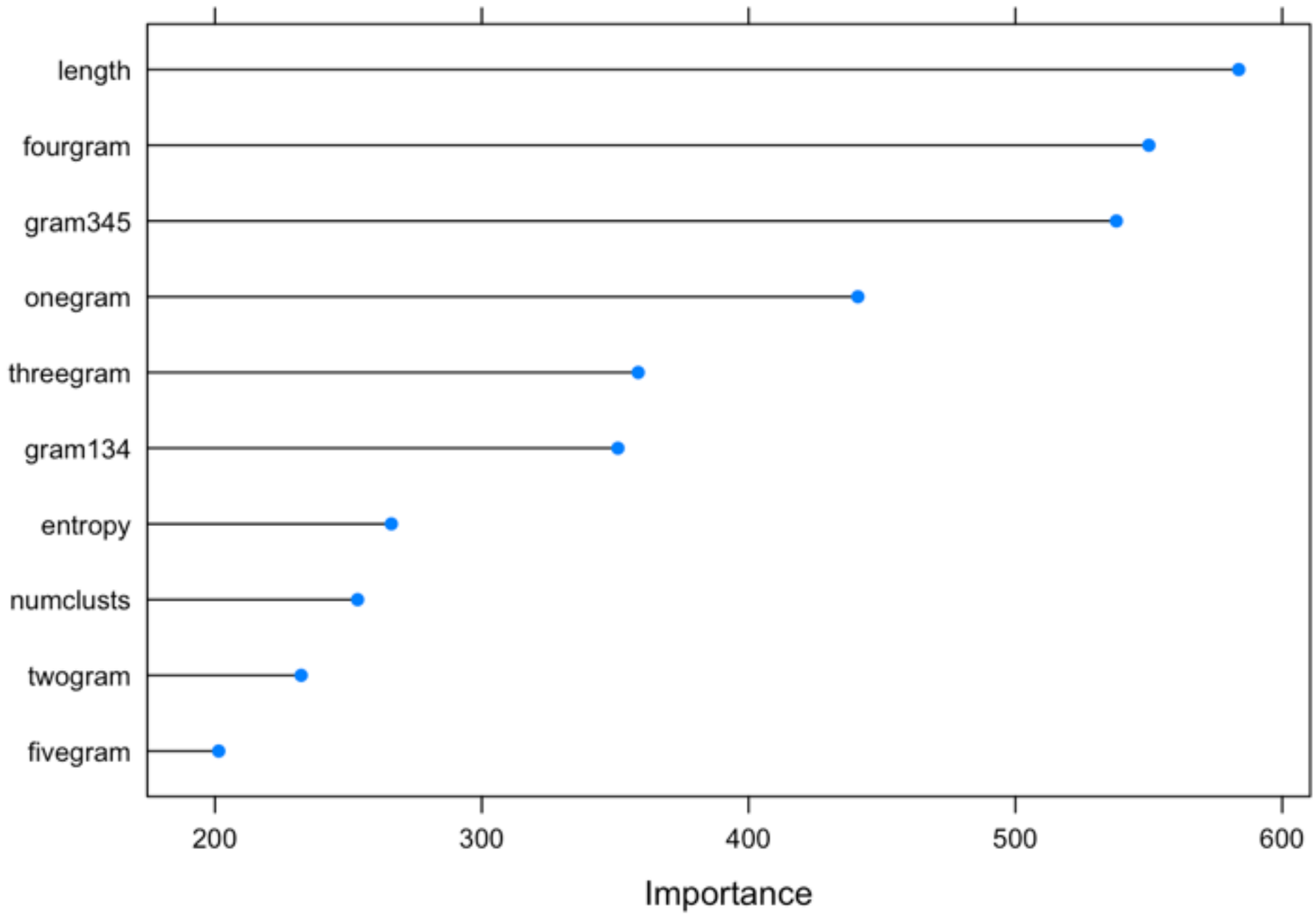


Python

Evaluating Features

Data Driven Security

```
rf.importance <- varImp(rfFit, scale=F)
plot(rf.importance)
```



Building a Classifier

Data Driven Security

1. Obtain labelled data
2. Add or derive features
3. Create test/train data sets
4. Run classification algorithm[s] on training data
5. Evaluate model and features on test data

More Evaluation

Data Driven Security

	dga	legit	domain
2	0.000	1.000	doubleclick
5	0.000	1.000	googlesyndication
6	0.000	1.000	googleapis
7	0.000	1.000	googleadservices
8	0.000	1.000	twitter
10	0.000	1.000	youtube
11	0.000	1.000	scorecardresearch
14	0.000	1.000	googleusercontent
17	0.006	0.994	msftncsi
22	0.000	1.000	verisign
24	0.000	1.000	quantserve
25	0.000	1.000	bluekai
31	0.000	1.000	digicert
34	0.000	1.000	pubmatic
36	0.000	1.000	adadvisor
43	0.006	0.994	yahooapis
47	0.000	1.000	googletagmanager
48	0.008	0.992	crwdcntrl
49	0.000	1.000	mookie1
54	0.000	1.000	imrworldwide

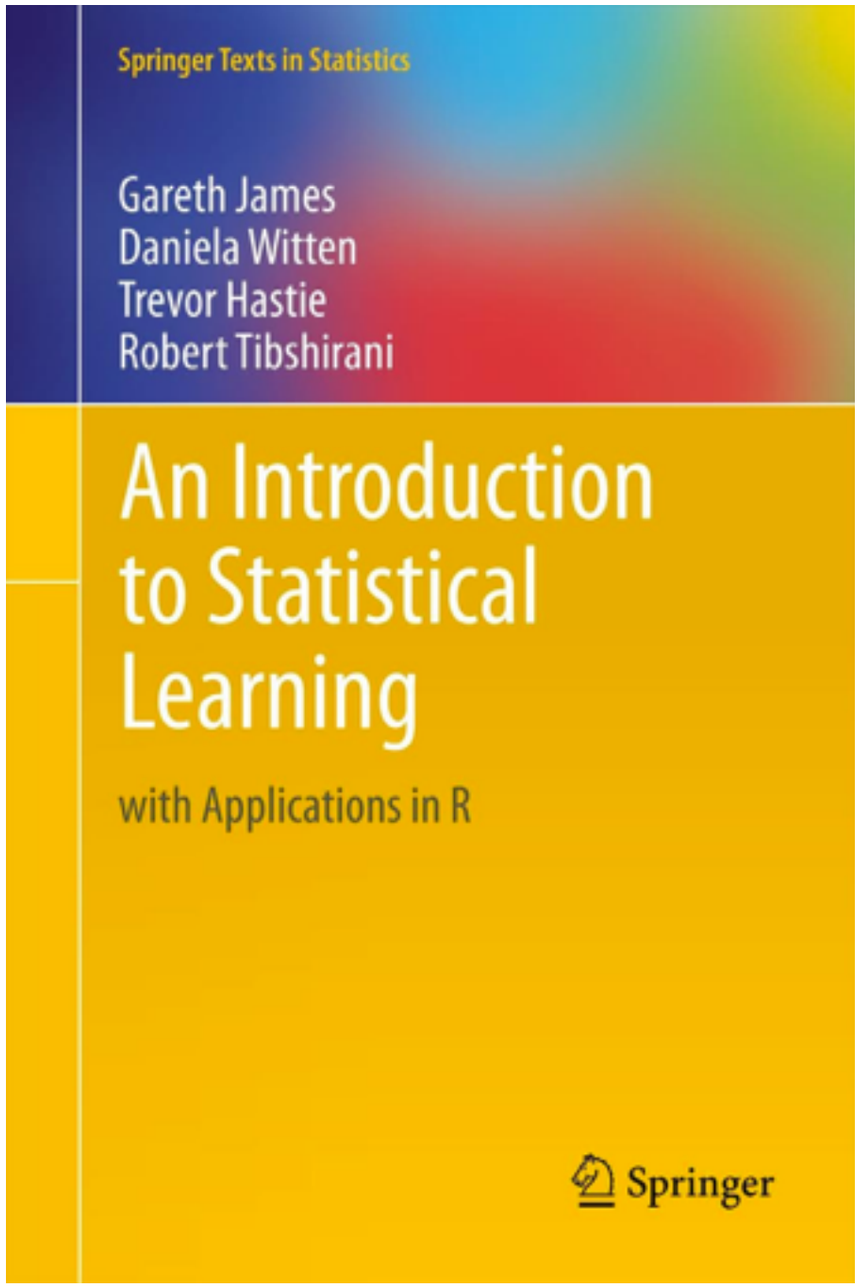
	dga	legit	domain
138957	1.000	0.000	7sy3v81toy7vim3br0410212pg
138958	1.000	0.000	i8hkuf1wwfc8w1g25u0110vx6w3
138959	1.000	0.000	etvp9c12ixta51jko7ba18xgd3
138961	1.000	0.000	bw25th1nsiukt1344bch1gwgr1h
138965	1.000	0.000	1opr1mm13rpbbm1iy7sdr1572kdu
138967	1.000	0.000	hhnp8p1732n9113wcdb2no89fb
138968	1.000	0.000	155xuit1i4td2bkc2t18qes6me
138969	1.000	0.000	5jndc1t1bvy811hk5ntxk6r4j
138971	1.000	0.000	p5b9an11o4kybhsghp2inlq58
138973	1.000	0.000	12sjxntztid4mh6snhldpqc3z
138974	0.998	0.002	15rrp3pyeoms11dbgsqurati8
138975	1.000	0.000	1wguzv3dd1tf9lwm6og2s6qkv
138976	1.000	0.000	1wvyjf21f8ve5967taqgpkpgvz
138977	1.000	0.000	r16k3i172flcb1u5d8vh1u7yfww
138978	1.000	0.000	1a3i2bq1cjka6s19kdymf1411282
138979	1.000	0.000	qcnqm211790taqp8h54eb9w85
138981	1.000	0.000	1ccvakyzxp80o1ij99er1d5yt56
138982	1.000	0.000	naihsdncxgv8e3eivnx2qmg0
138983	1.000	0.000	5yeo3813dmjen17zzsxtp1oyhn
138984	0.992	0.008	1gl0zqzcqo5futz9ead2vj6e3

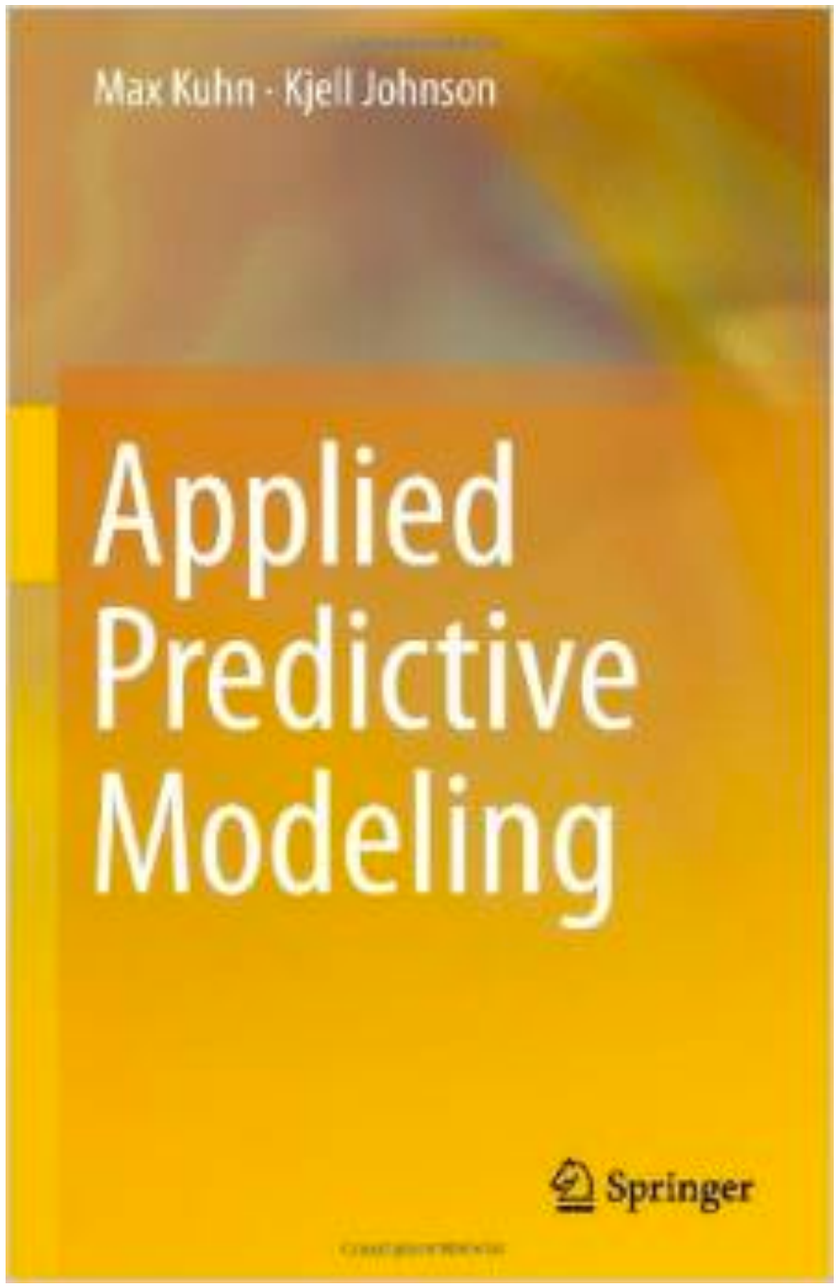
	dga	legit	domain
96375	0.532	0.468	muskelschmiede
96739	0.492	0.508	cendrawasih11
97182	0.506	0.494	empayar-pemuda
97824	0.506	0.494	avto-flagman
26011	0.534	0.466	semilukskaya-crb
25273	0.502	0.498	amovpnforoosh11
27955	0.482	0.518	fairheadkenya
3356	0.536	0.464	m3mieszkania
35484	0.524	0.476	stukadoorsbedrijfvannoord
3876	0.504	0.496	pik-equipment
41173	0.520	0.480	**oxfordlawtrove**
71022	0.546	0.454	inezandvinoodh
72228	0.528	0.472	**voiceofdaegu**
99001	0.536	0.464	sacdokulmesi-tr
878461	0.452	0.548	viokbmsinerce
878951	0.512	0.488	hebsphsplitih
886501	0.504	0.496	hotodfonwpougi
890121	0.544	0.456	vgcjamateqgut
897231	0.504	0.496	bjoseraicgty
912801	0.470	0.530	ewebqestbocrus
916521	0.496	0.504	dseemngarkpll
919051	0.528	0.472	ojmeatrfojec
924721	0.478	0.522	**plletrsadpawy**
936541	0.528	0.472	fbyantymmandfxh
93780	0.514	0.486	yreapptcabyech

Evaluating visually

Data Driven
Security

Recommendations





Jay Jacobs
@jayjacobs