RSAConference2015
San Francisco | April 20-24 | Moscone Center

SESSION ID: ANF-T07R

Security Data Science: From Theory to Reality

Jay Jacobs

Security Data Scientist
Verizon Security Research
@jayjacobs

Bob Rudis

Security Data Scientist
Verizon Security Research
@hrbrmstr

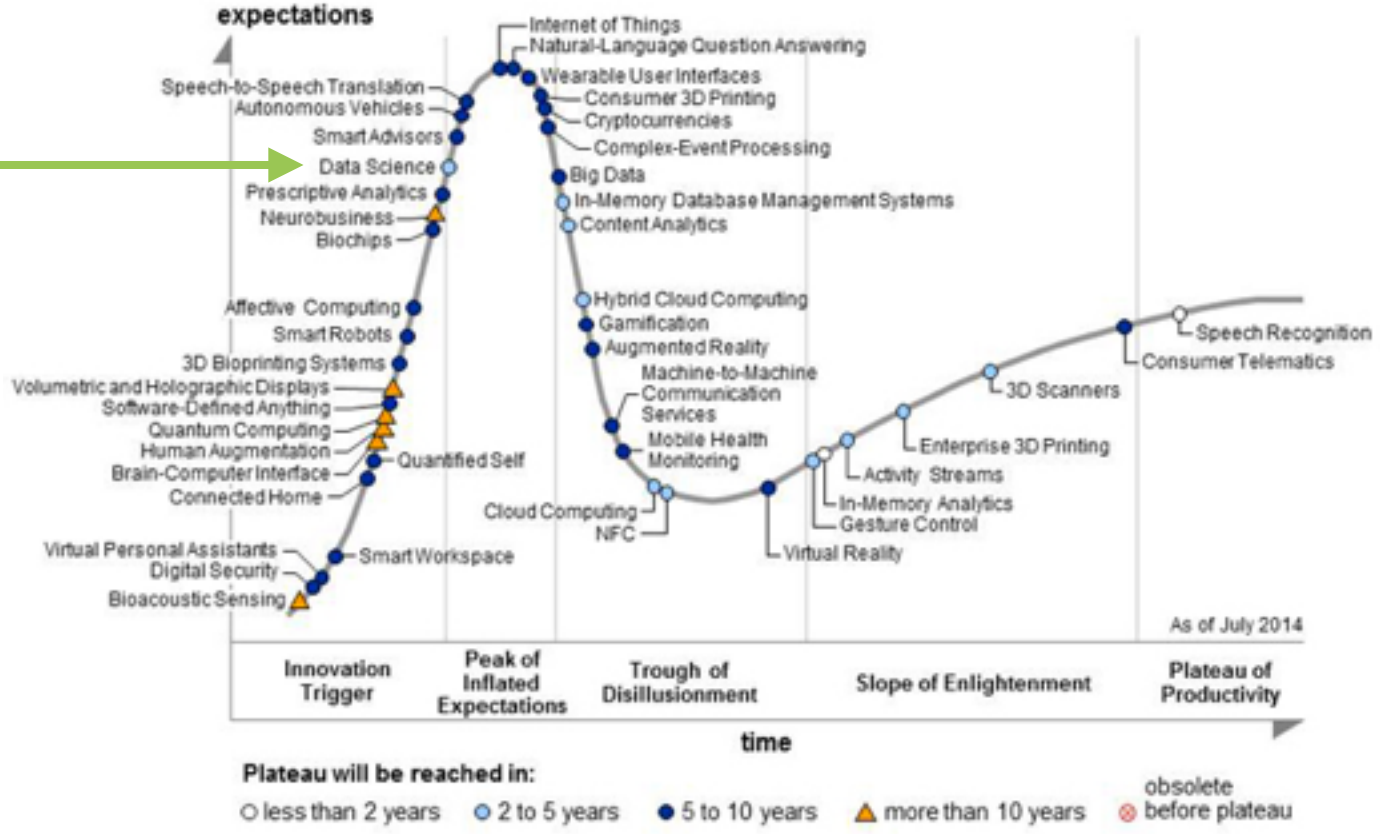
DBIR: http://www.verizonenterprise.com/DBIR/
Book: http://dds.ec/amzn
AOL Keyword: DBIR

Blog: http://datadrivensecurity.info/blog
Podcast: http://datadrivensecurity.info/podcast

@ddsecblog • @ddsecpodcast
@jayjacobs • @hrbrmstr

[Security] Data Science

We are here



SOURCE: Gartner, August 2014

Data Science is…

- "Data scientist is just a sexed up word for statistician." *- Nate Silver*
- Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician. *- Josh Wills, Ex-Statistician, Data Scientist at Cloudera*
- “Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.” - *Jeff Leek, JHU/Coursera*

[Security] Data Science is…

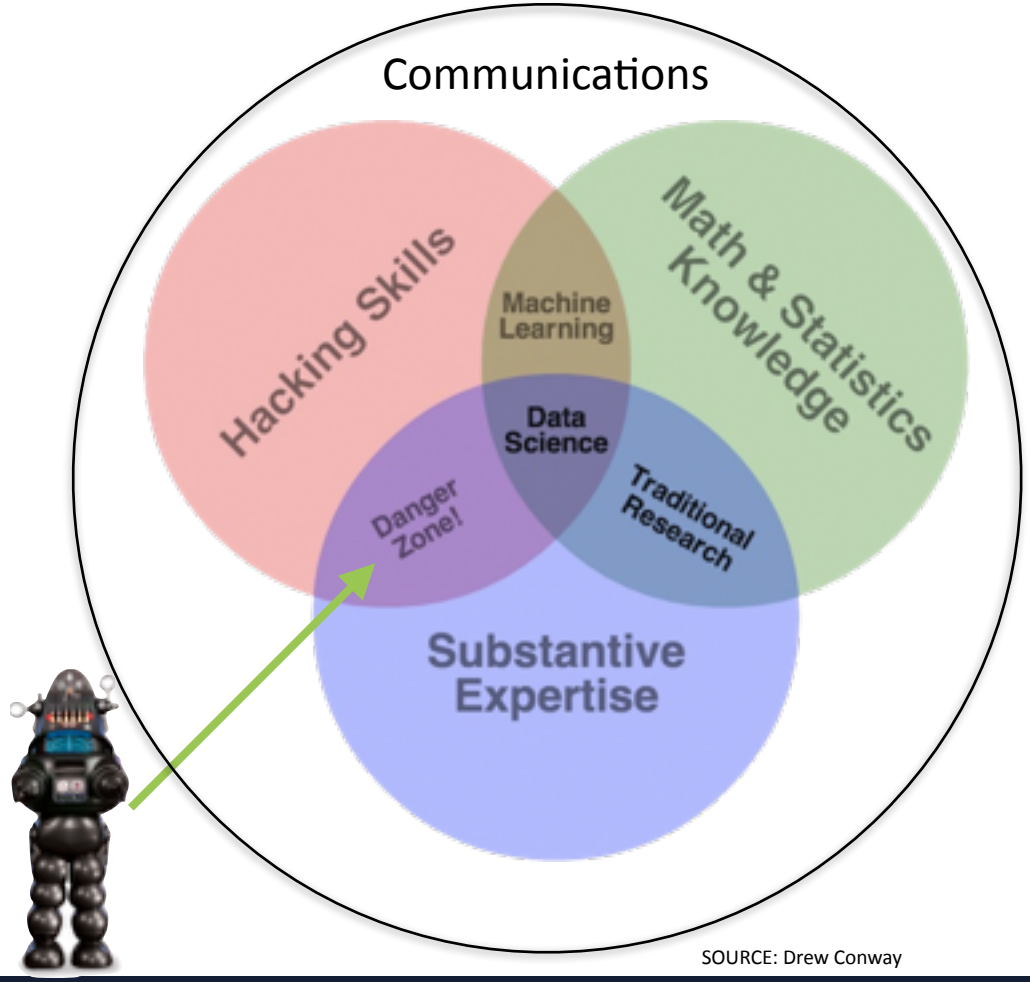
a shift from security as opinions and blind "best practice" towards **security as a science**.

- Are insiders more of a threat that external actors?
- Should new patches take precedence over old?
- Where should I invest my security budget?
- (question the stuff that can actually make a difference)

SOURCE: Drew Conway

Basic Process

- Form a [Research] Question
- Acquire & “clean” data
- Analyze Data
- Examine Outcomes
- Visualize & Communicate Results
- Lather, rinse, repeat



SOURCE: Drew Conway

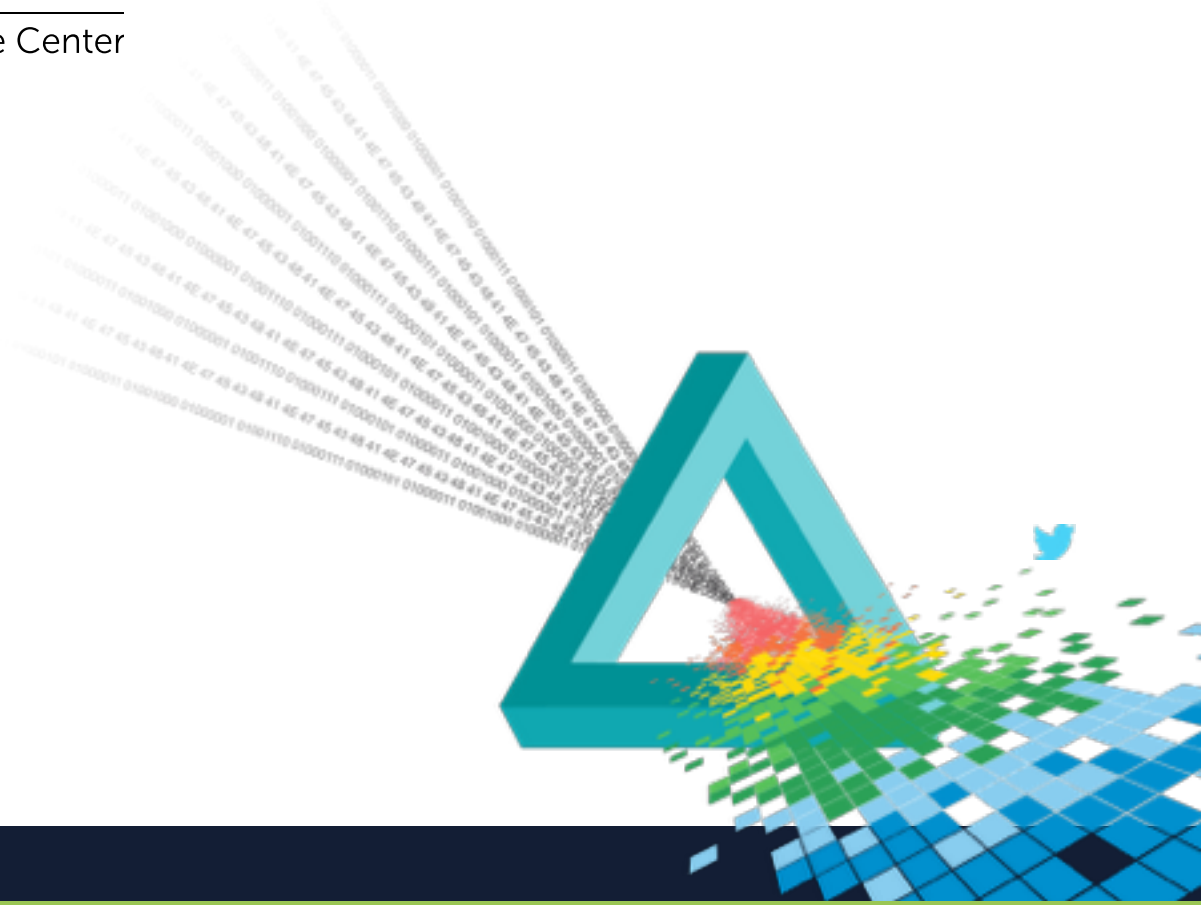
Danger Zone Process

- Get some (any/convenient) data
- COUNT ALL THE THINGS
- Make a dashboard / PowerPoint

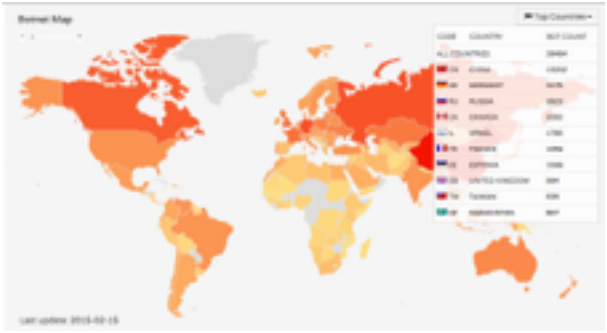
RSAConference2015

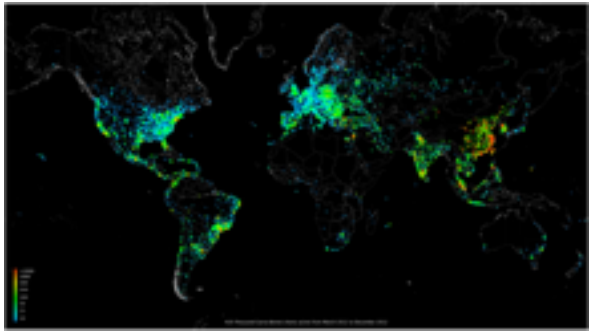
San Francisco | April 20-24 | Moscone Center

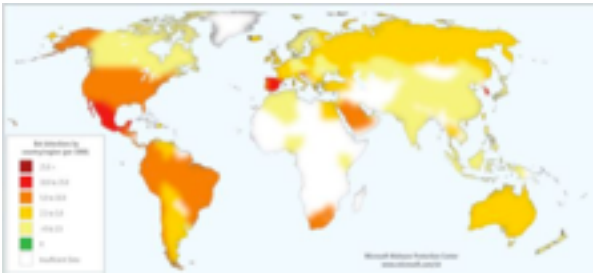
Finding Your Way In IPv4 Space

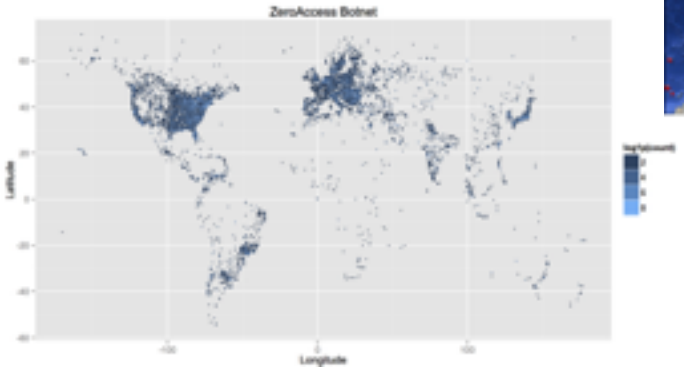



C&C








ZeroAccess Botnet
Latitude
Longitude

Representation of IPv4 Space

- IPv4 address
  - 32-bit integer canonically represented by 4 octets (“`10.20.30.40`”)
  - Fits inside a subnet (“`10.20.30.0/24`”) which is nothing more than a range of 32-bit integers
  - **We can use this to come up with a better way of assigning “latitude” and “longitude”**
- On the internet, IPv4 blocks are allocated to regional registries, grouped into Autonomous System (AS) numbers
  - These registries then assign AS numbers to organizations
  - **We can use this instead of (or along with) “country” & “city”**

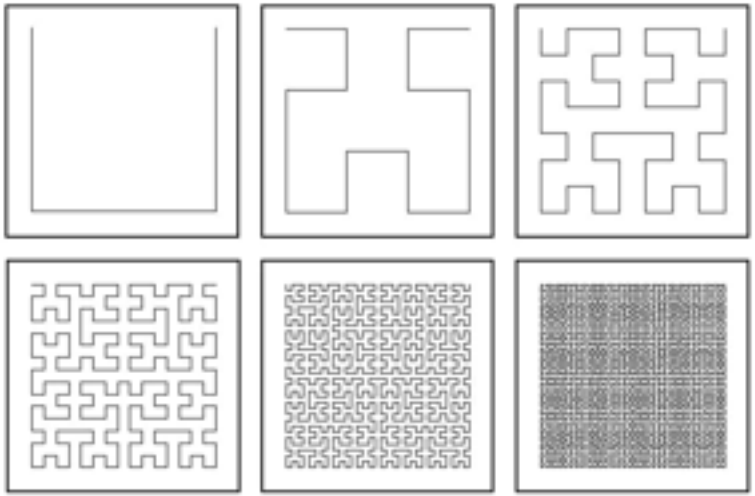
But first, we need some old-school math

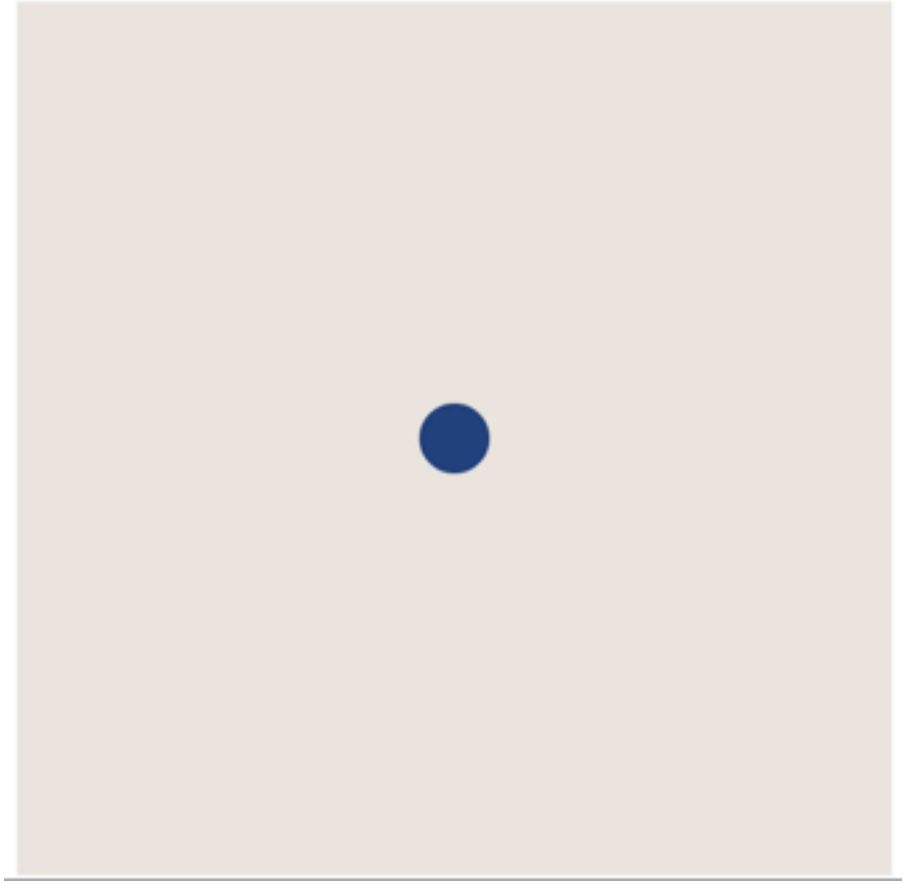
Visual Representation of IPv4 Space

1891

David Hilbert

Hilbert Curves

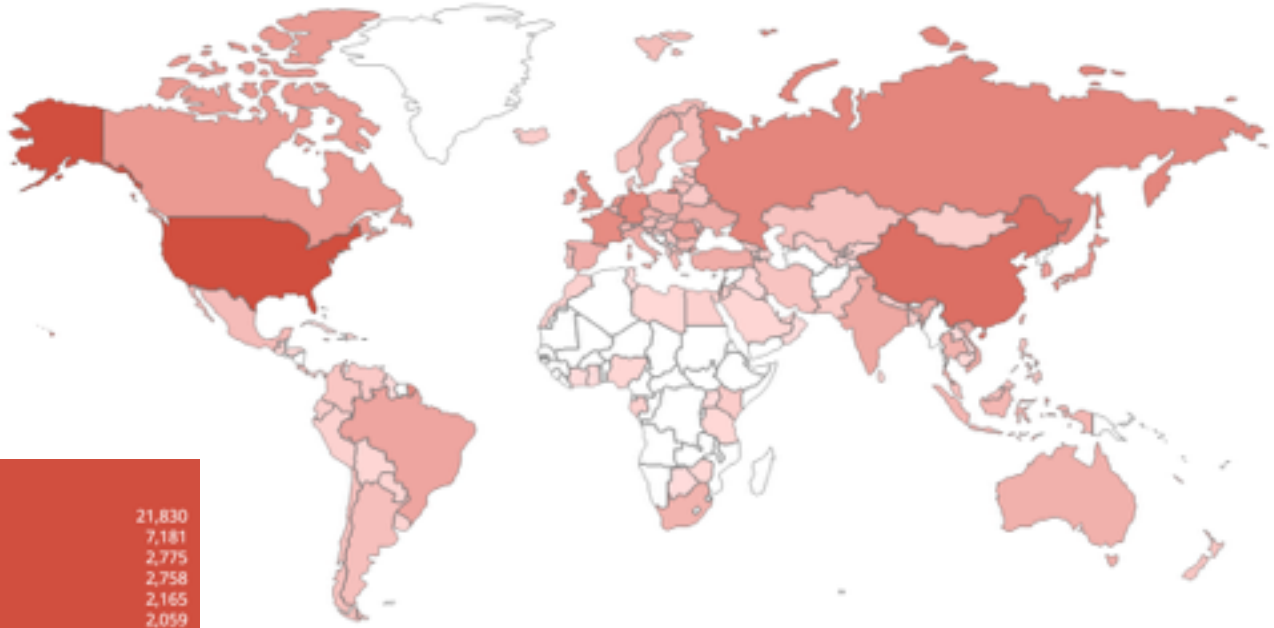




Visual Representation of IPv4 Space

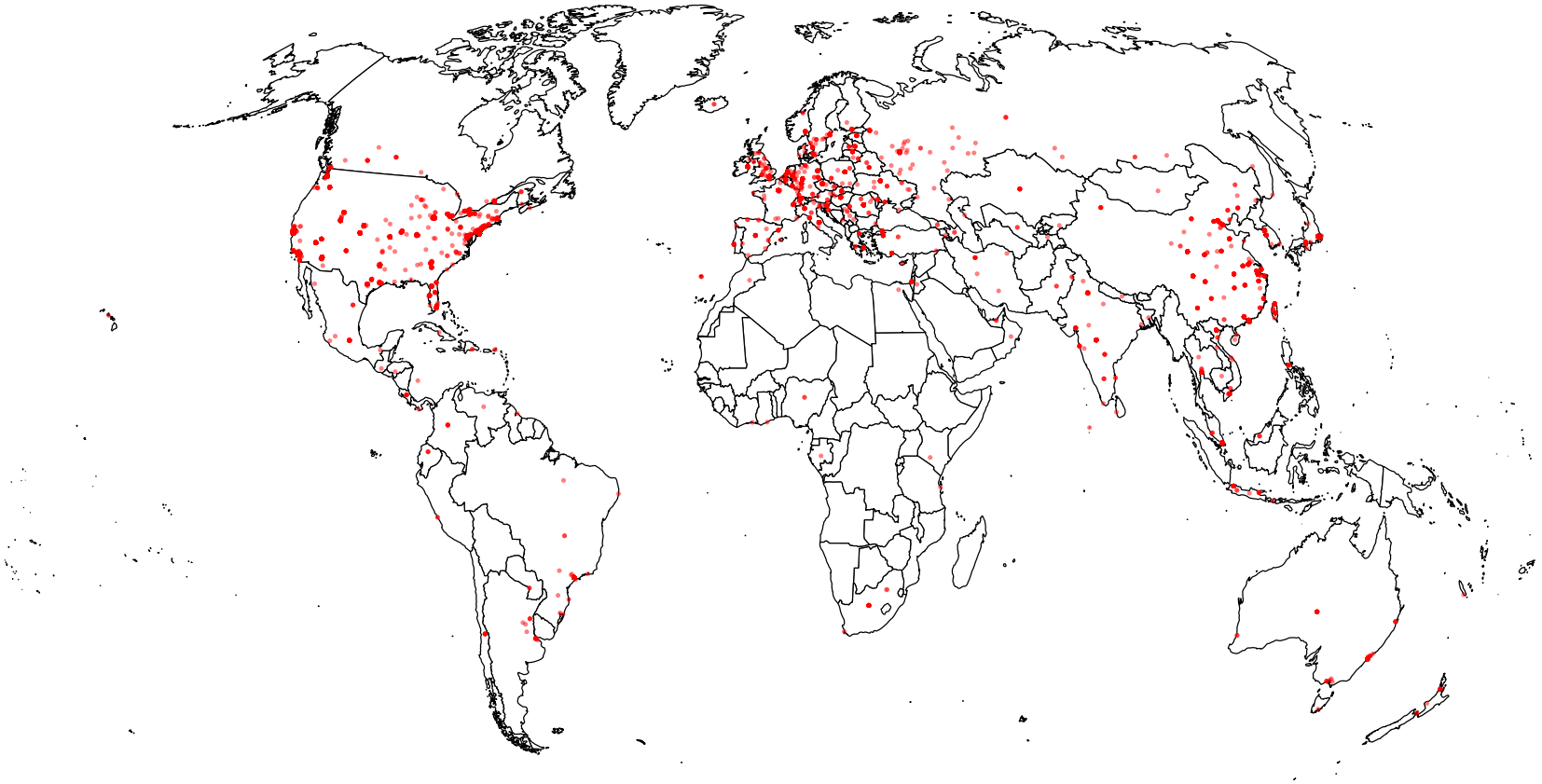
1891

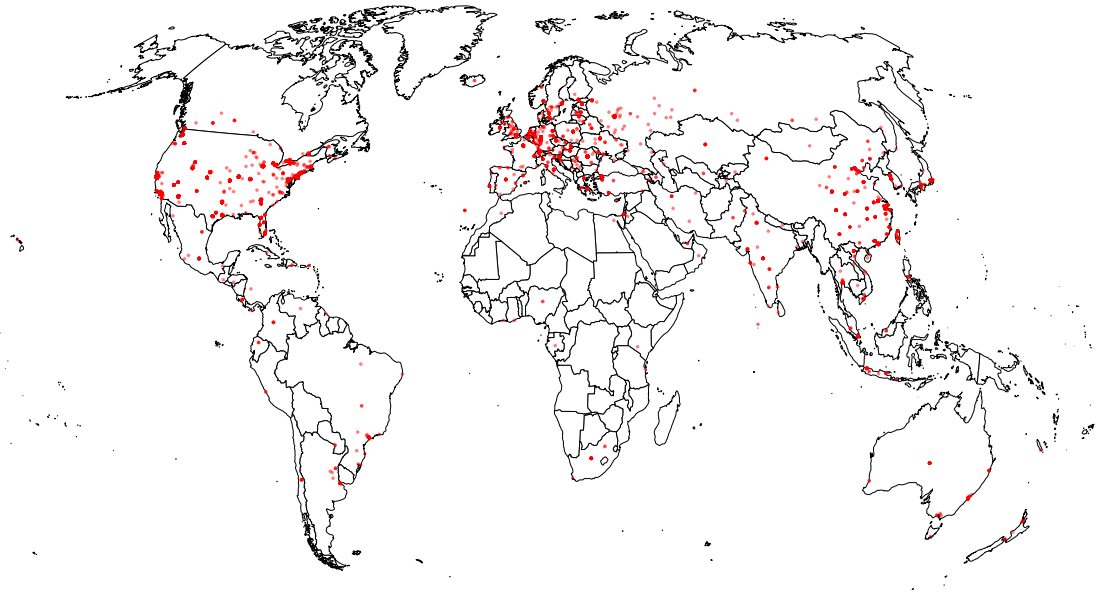
2006

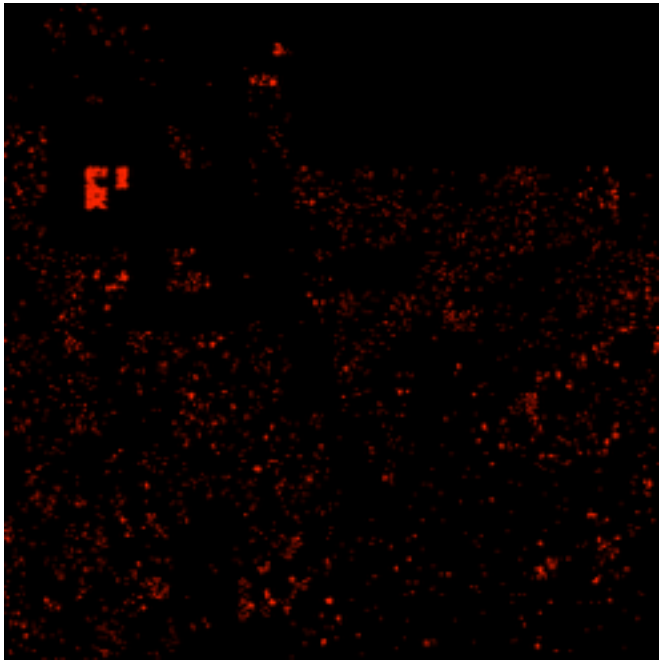


Top Countries	
1. United States	21,830
2. China	7,181
3. Russian Federation	2,775
4. Germany	2,758
5. France	2,165
6. United Kingdom	2,059
7. Japan	1,327
8. Netherlands	1,304
9. Canada	997
10. Romania	920

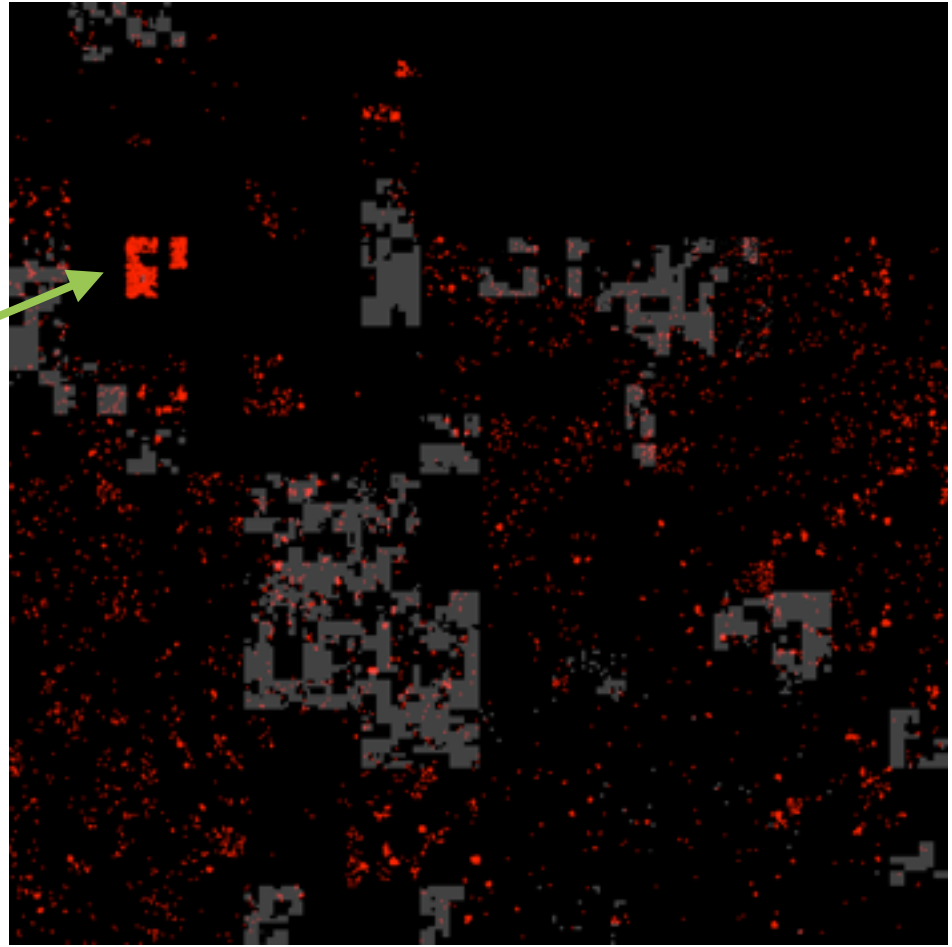
https://www.shodan.io/report/YQ8ayHWi

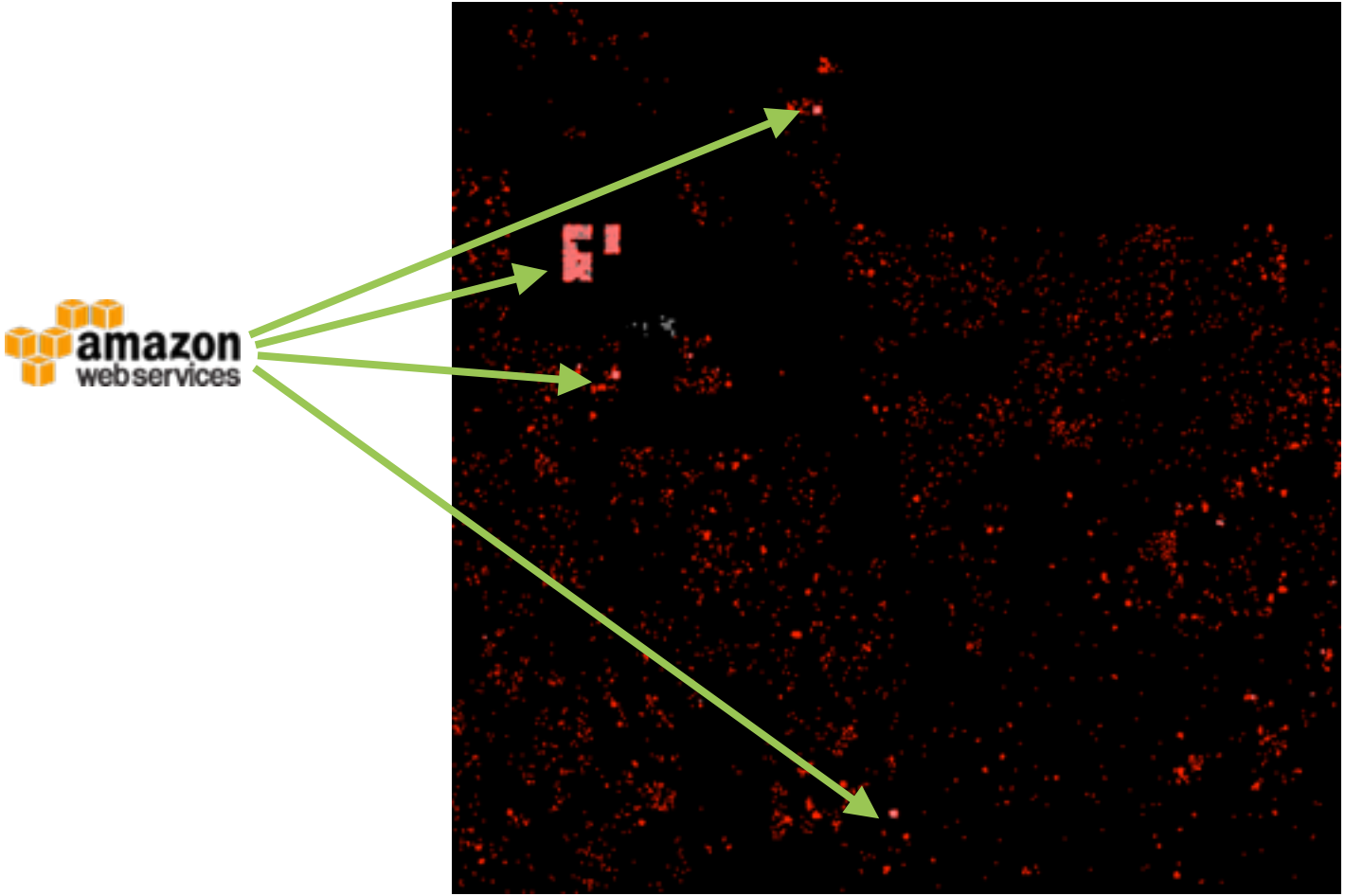


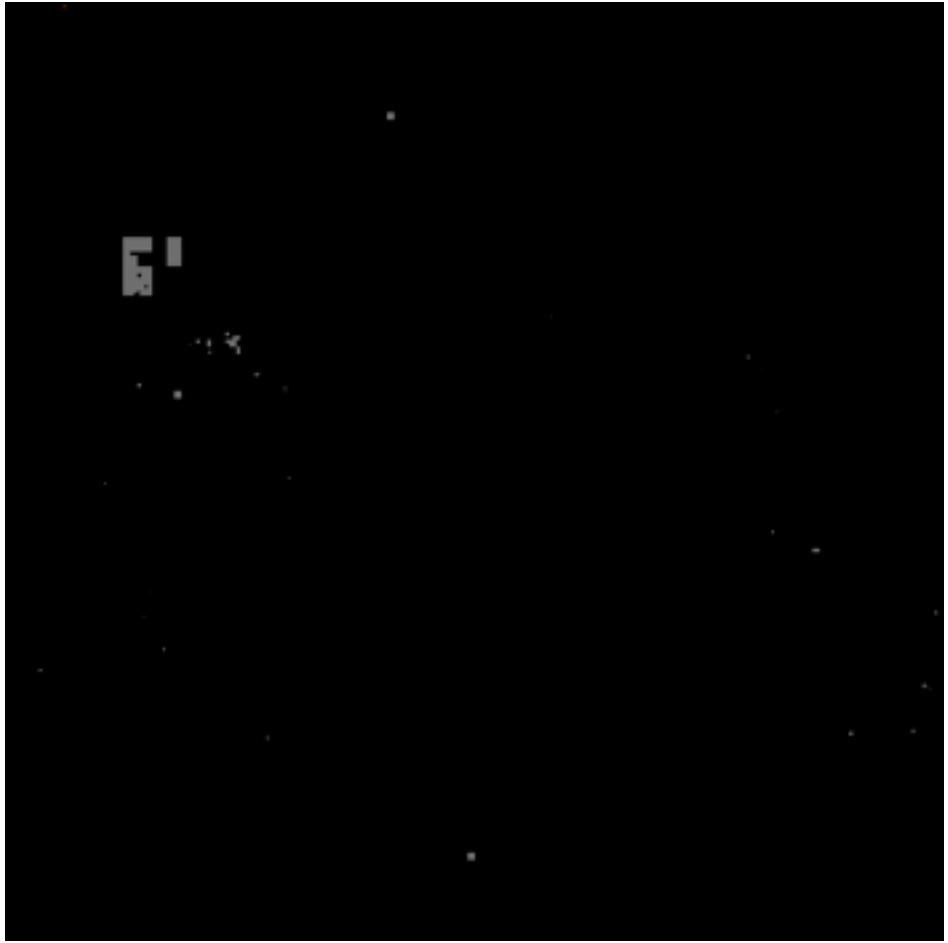


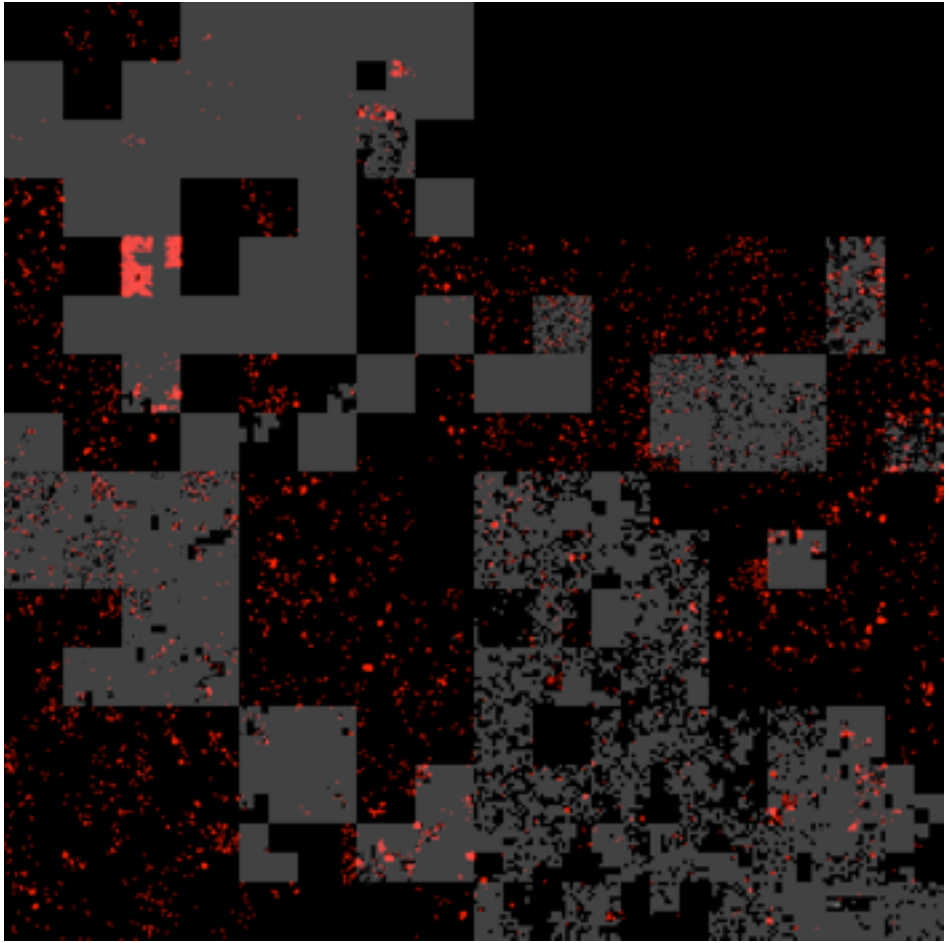


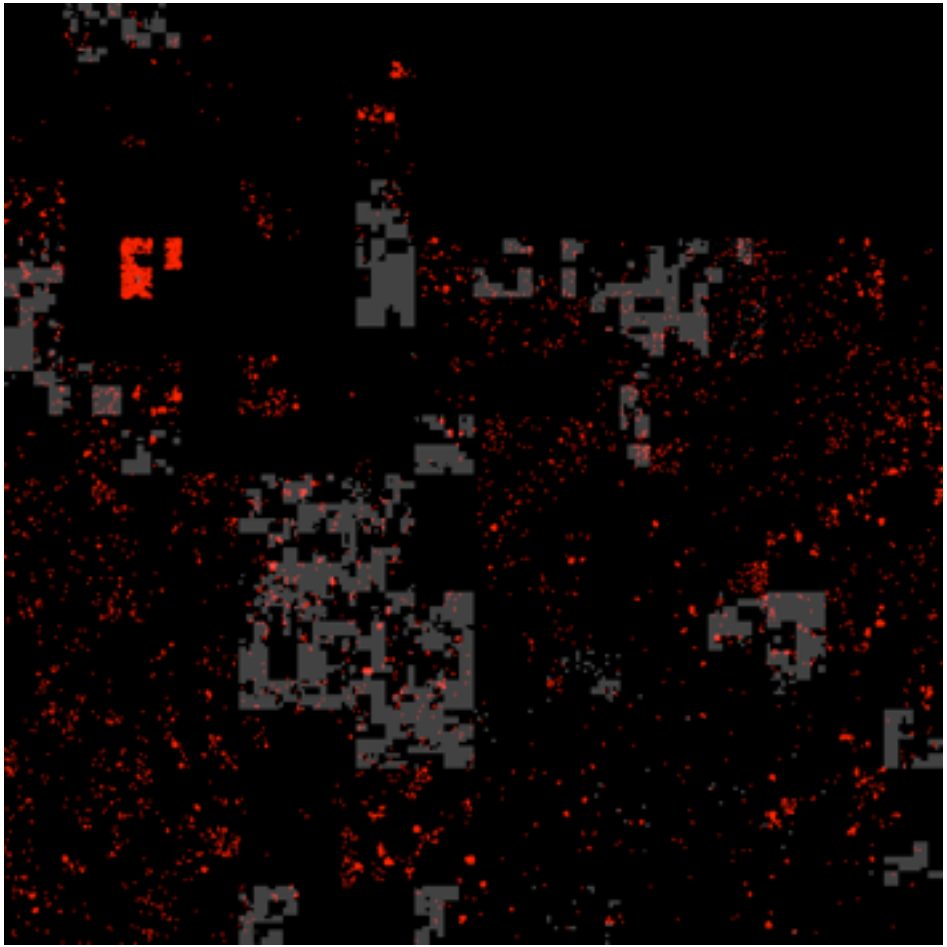
But, what's this?

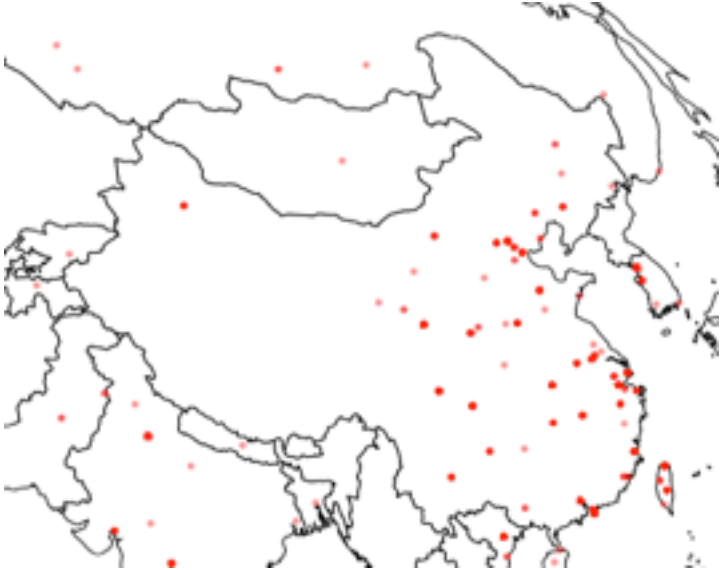


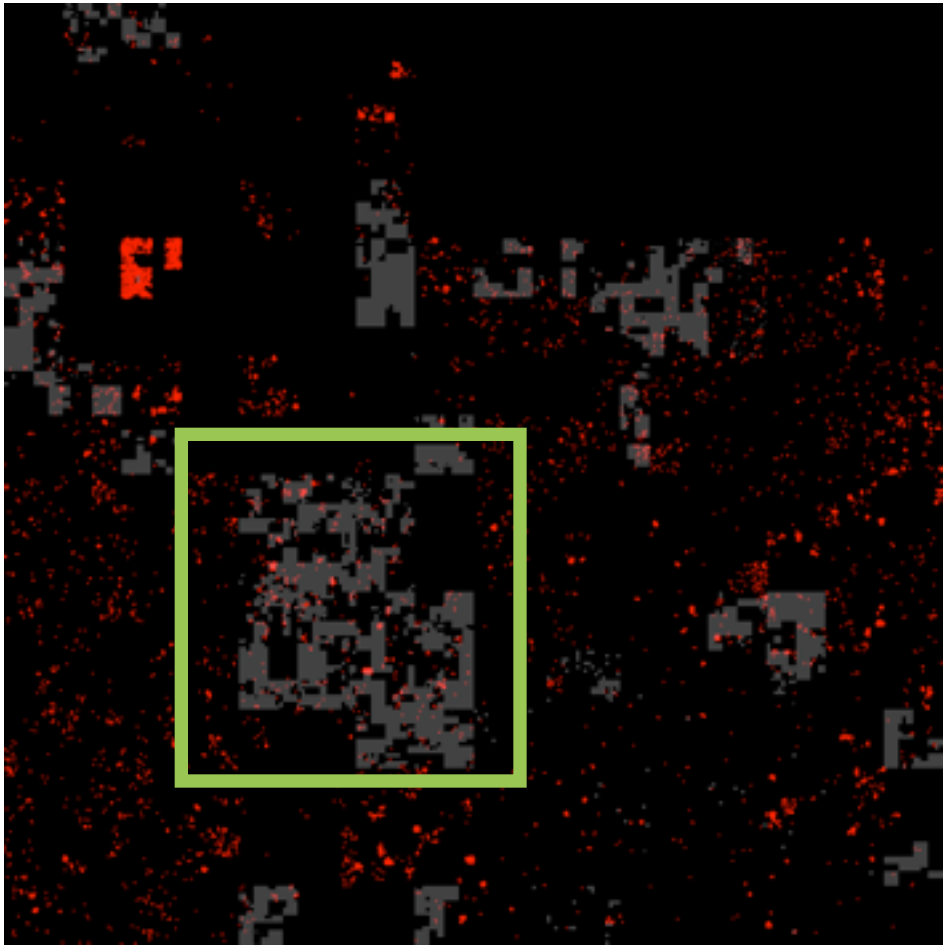

amazon
web services

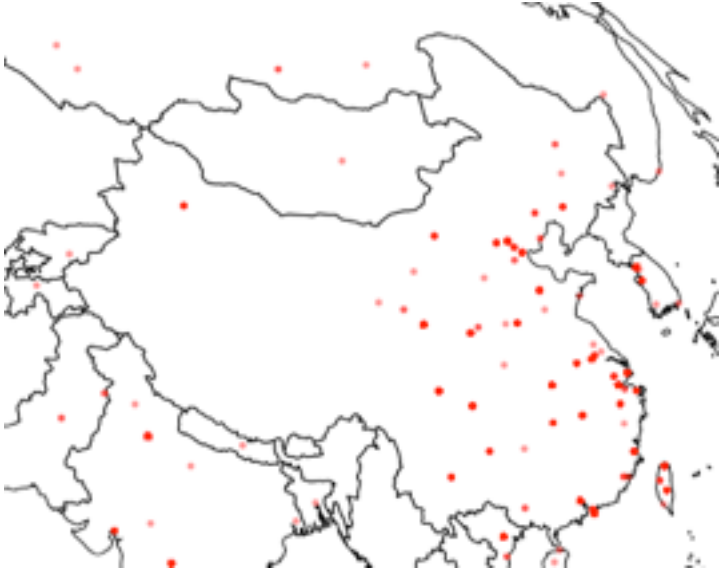


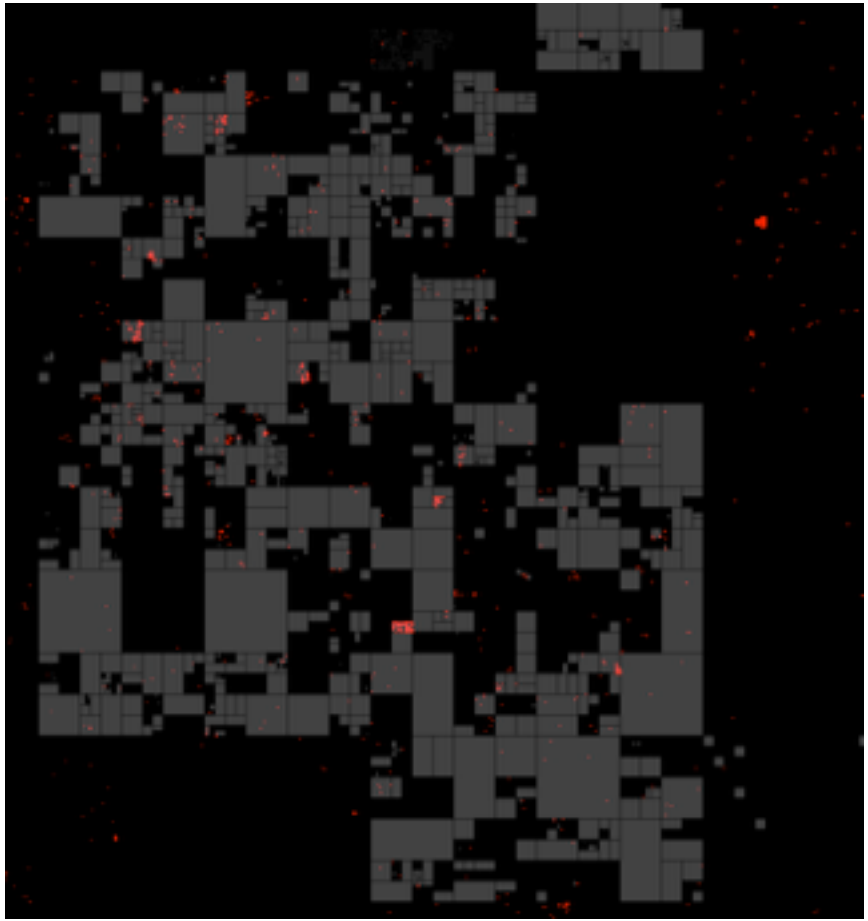


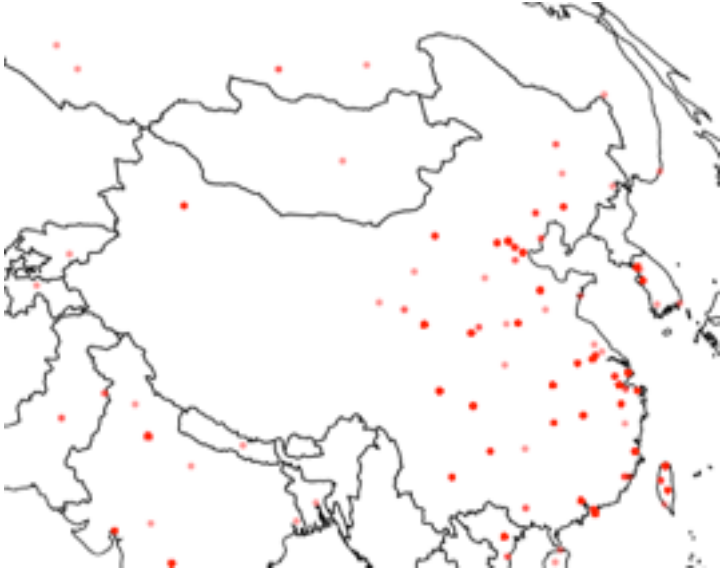


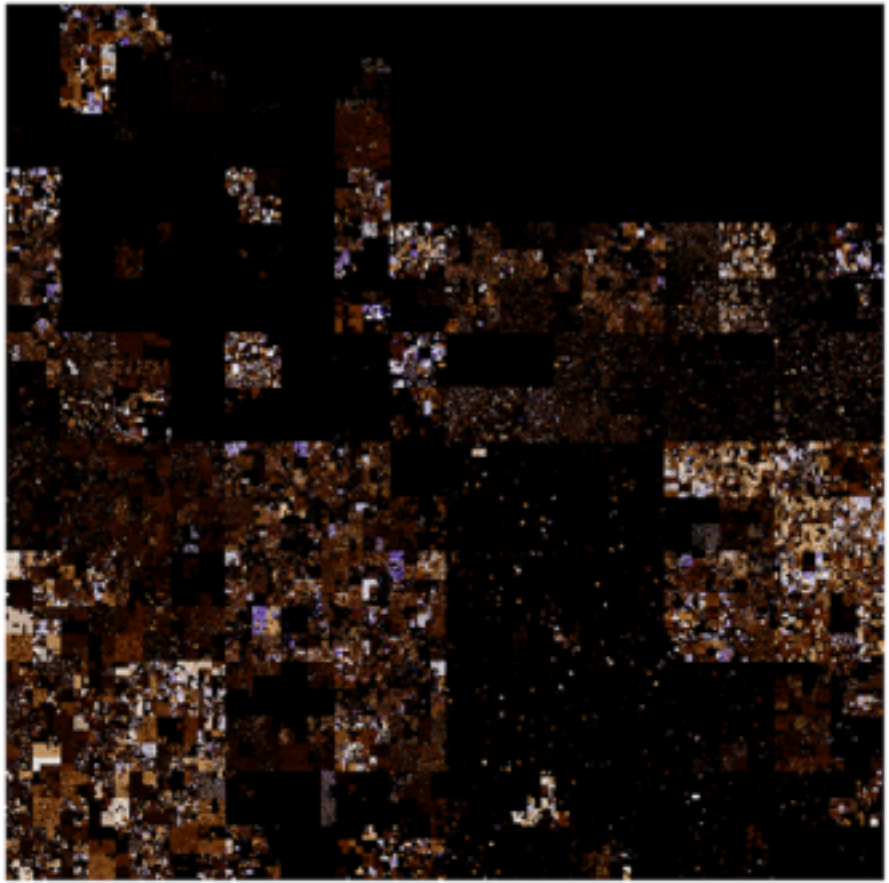




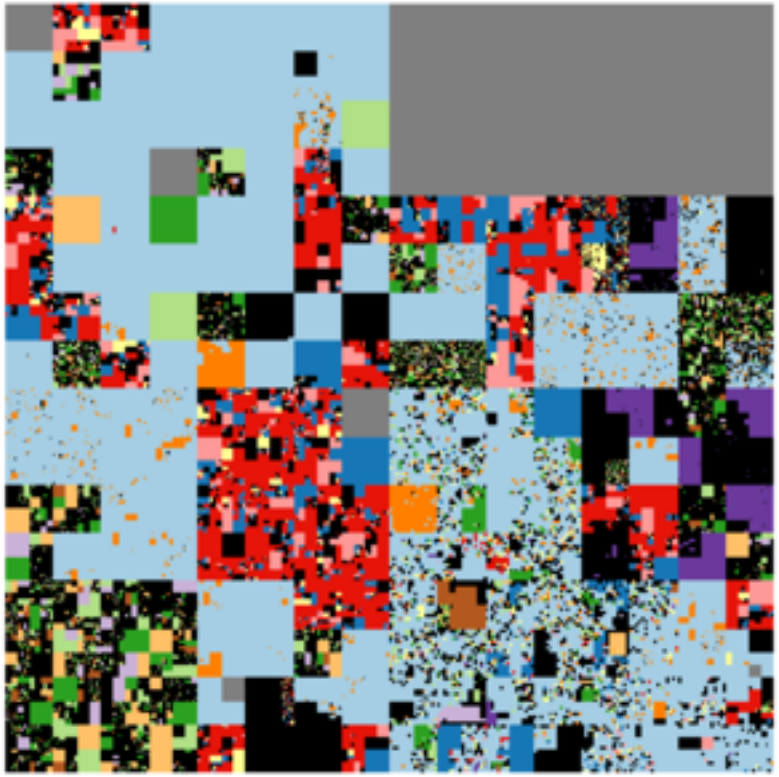








Top 12 Countries of the Internet



http://bit.ly/ipv4hilvis

So What?

- Visual cue for unusual hotspots
- Get a handle on the home front
- Impress your colleagues by saying “12th-order Hilbert curve”

TRY THIS AT HOME!

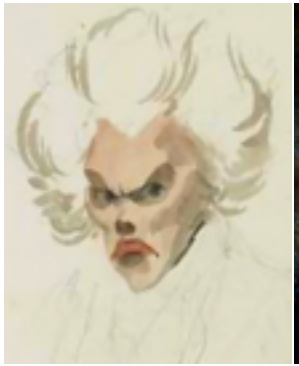
- https://github.com/vz-risk/ipv4heatmap
- http://maps.measurement-factory.com/software/ipv4-heatmap.1.html

RSAConference2015

San Francisco | April 20-24 | Moscone Center

Insight from Regression

Least Squares to Ponemon



Adrien-Marie Legendre
(1805)

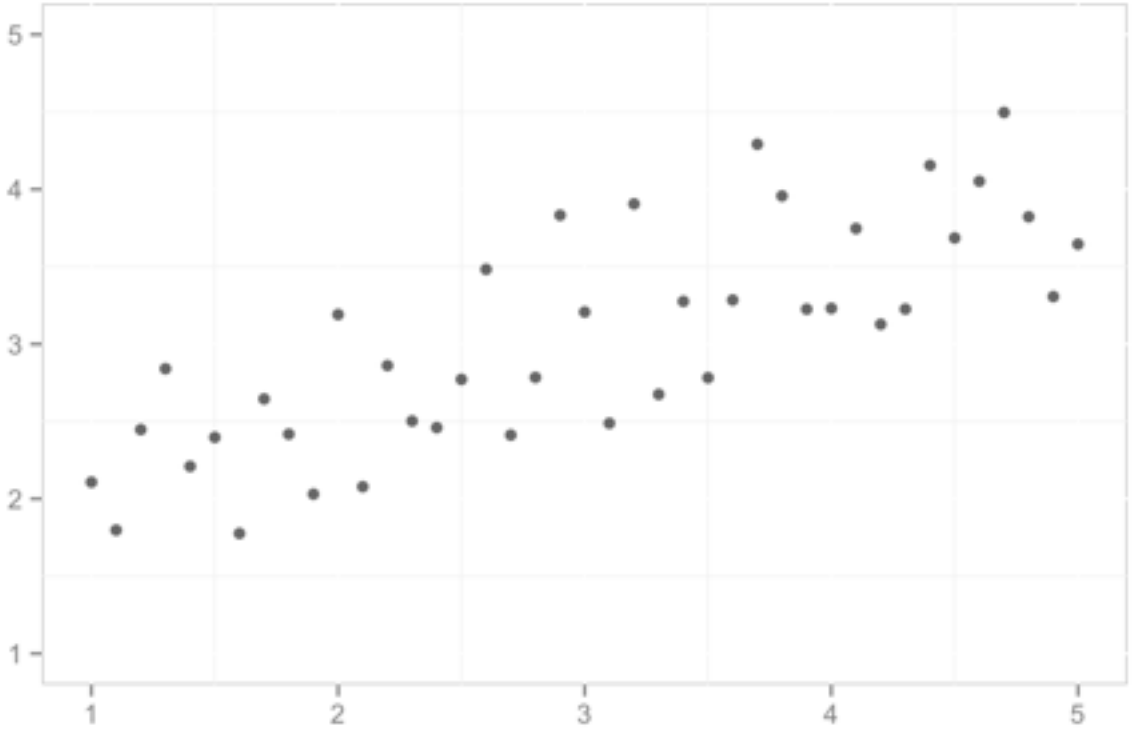
Carl Frederic
Gauss (1809)

Francis Galton
(1899)

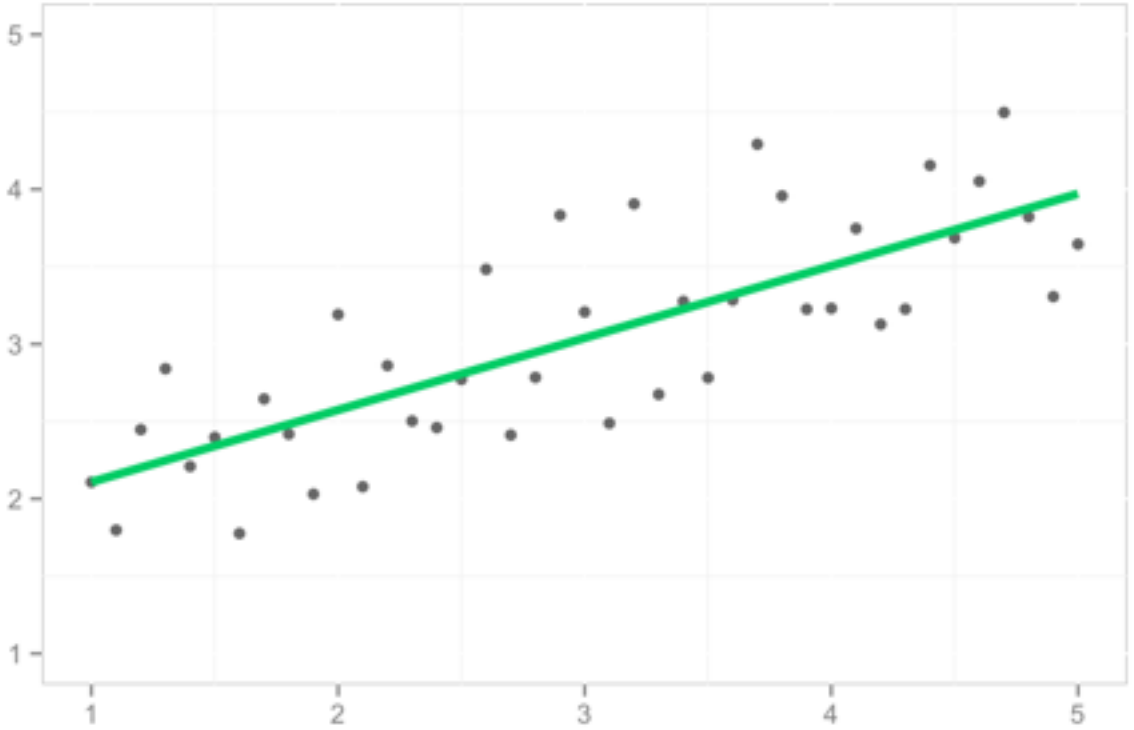
Karl Pearson
(1903)

Ronald A. Fisher
(1922)

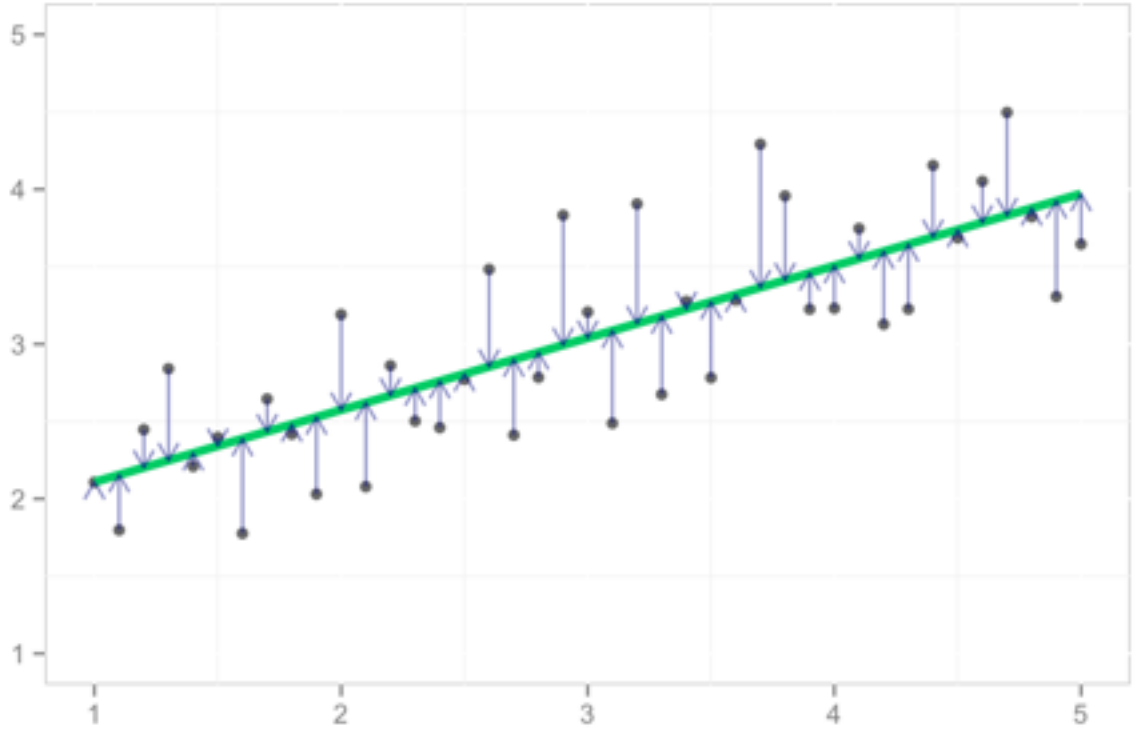
Least Squares to Ponemon



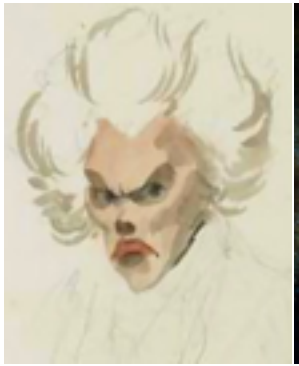
Least Squares to Ponemon



Least Squares to Ponemon

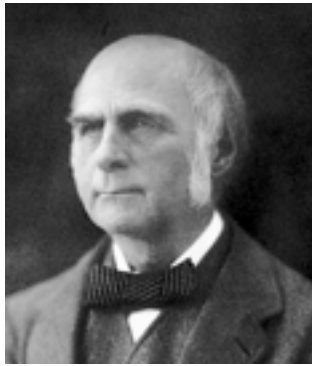


Least Squares to Ponemon



Adrien-Marie Legendre
(1805)

Carl Frederic
Gauss (1809)

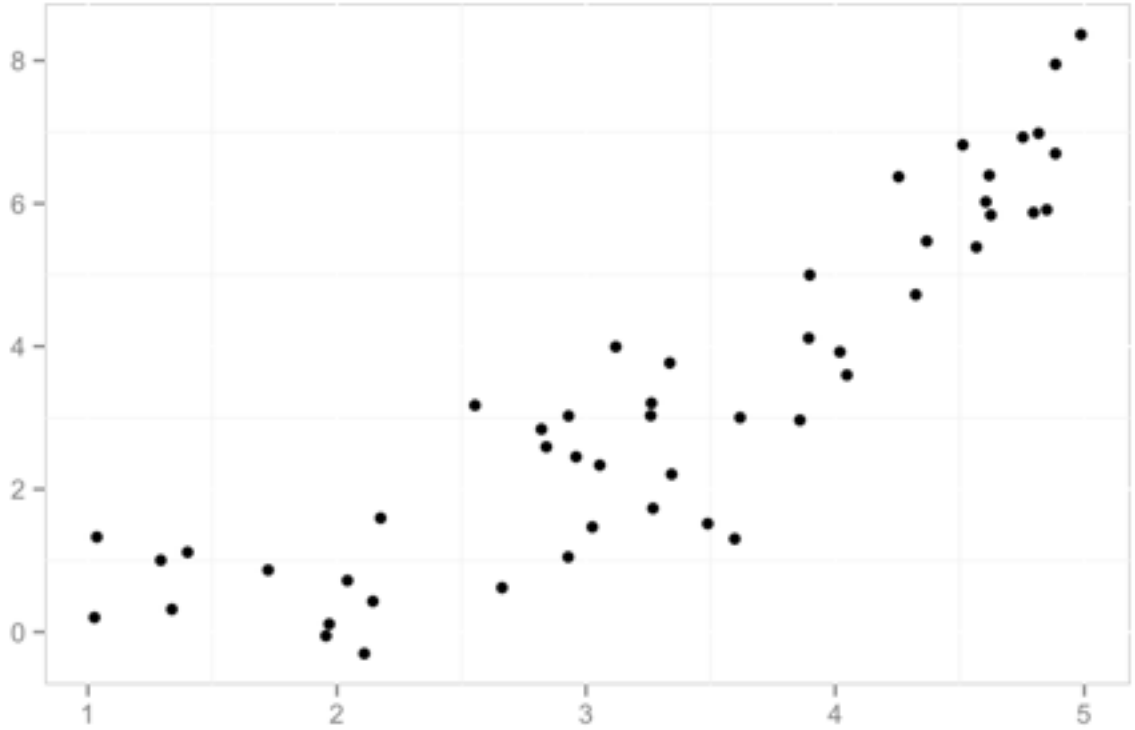


Francis Galton
(1869)

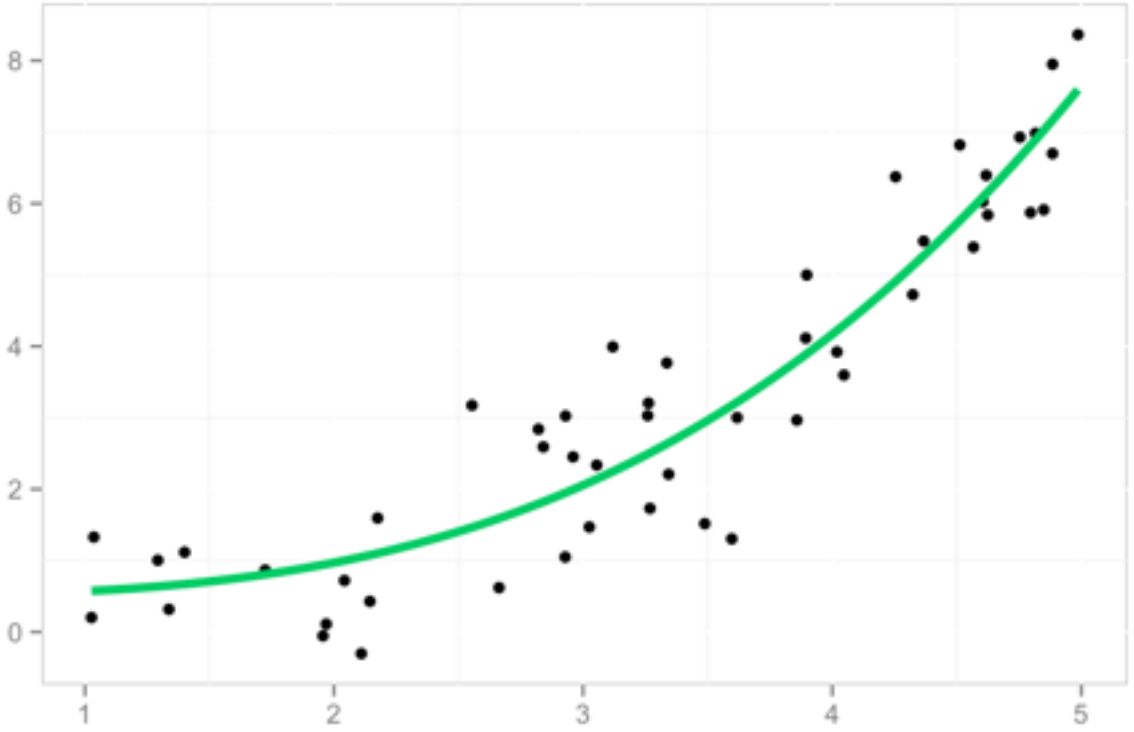
Karl Pearson
(1903)

Ronald A. Fisher
(1922)

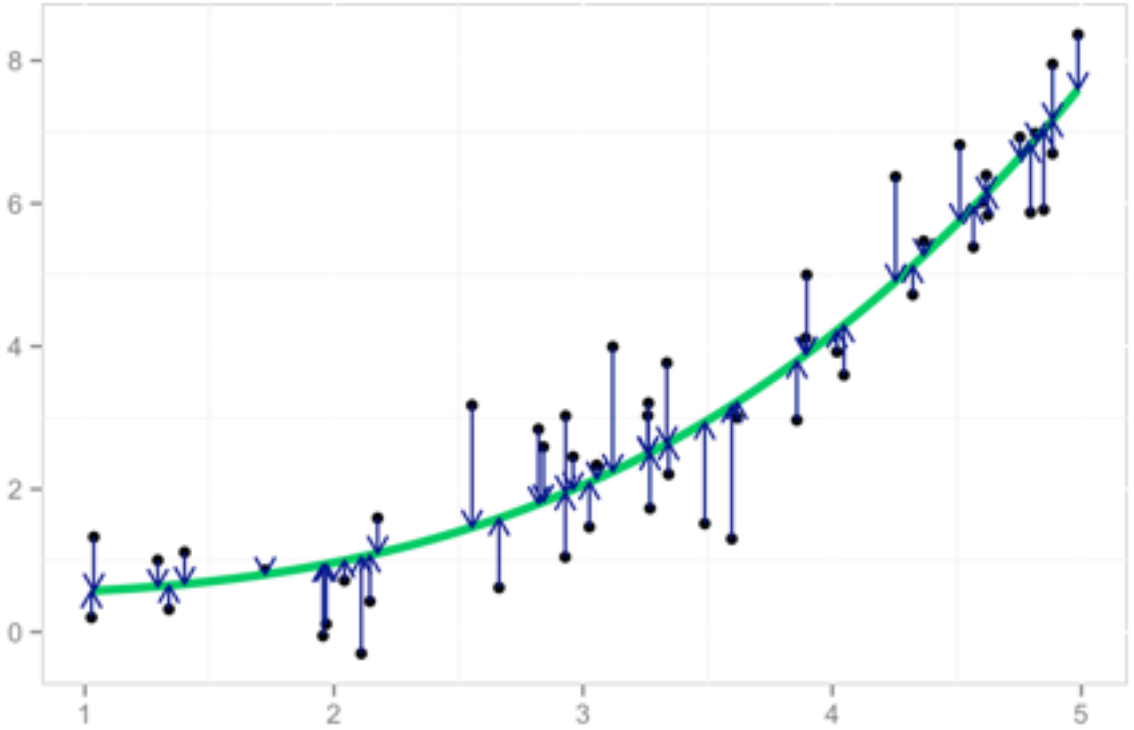
Least Squares to Ponemon



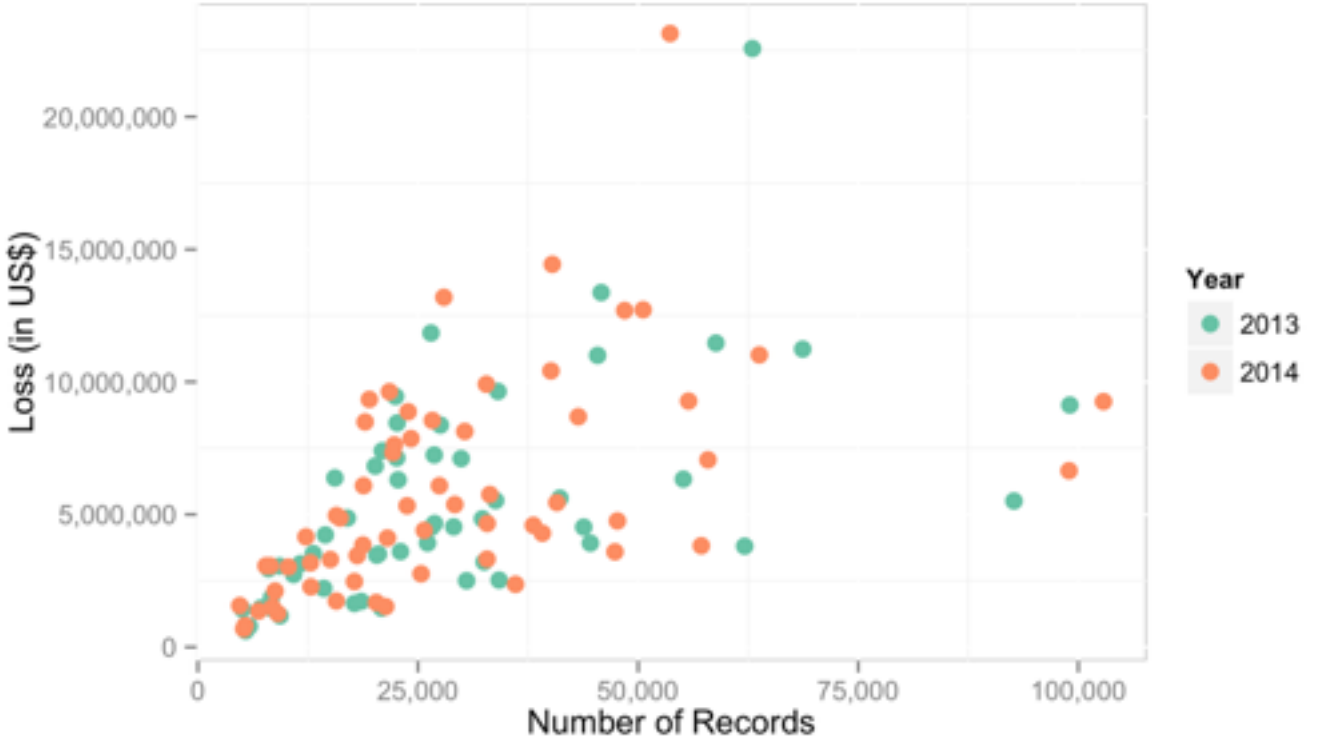
Least Squares to Ponemon



Least Squares to Ponemon



Least Squares to Ponemon

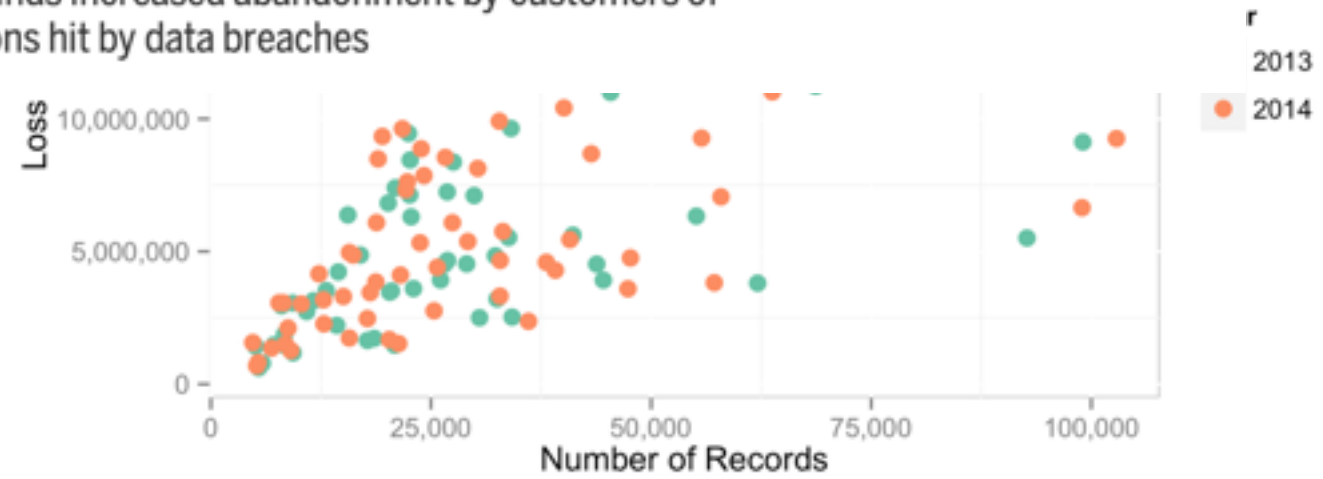


NETWORKWORLD Most read:

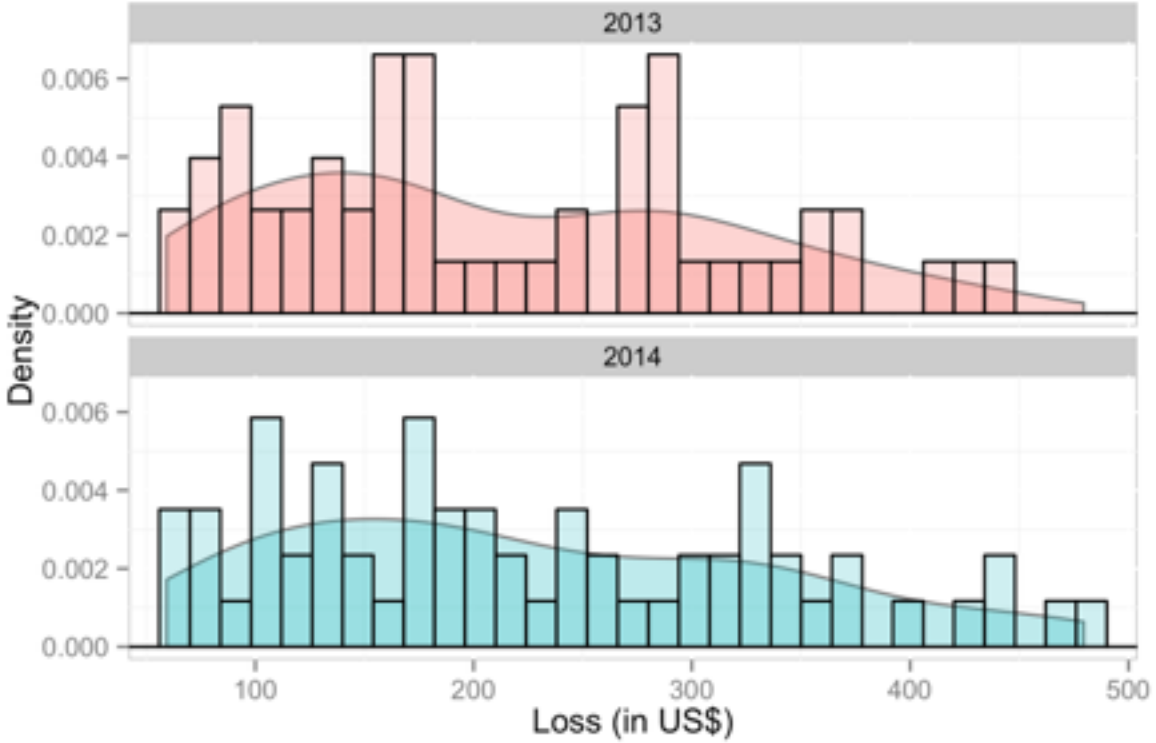
Home > Security > Malware/Cybercrime

Data breaches 9% more costly in 2013 than year before

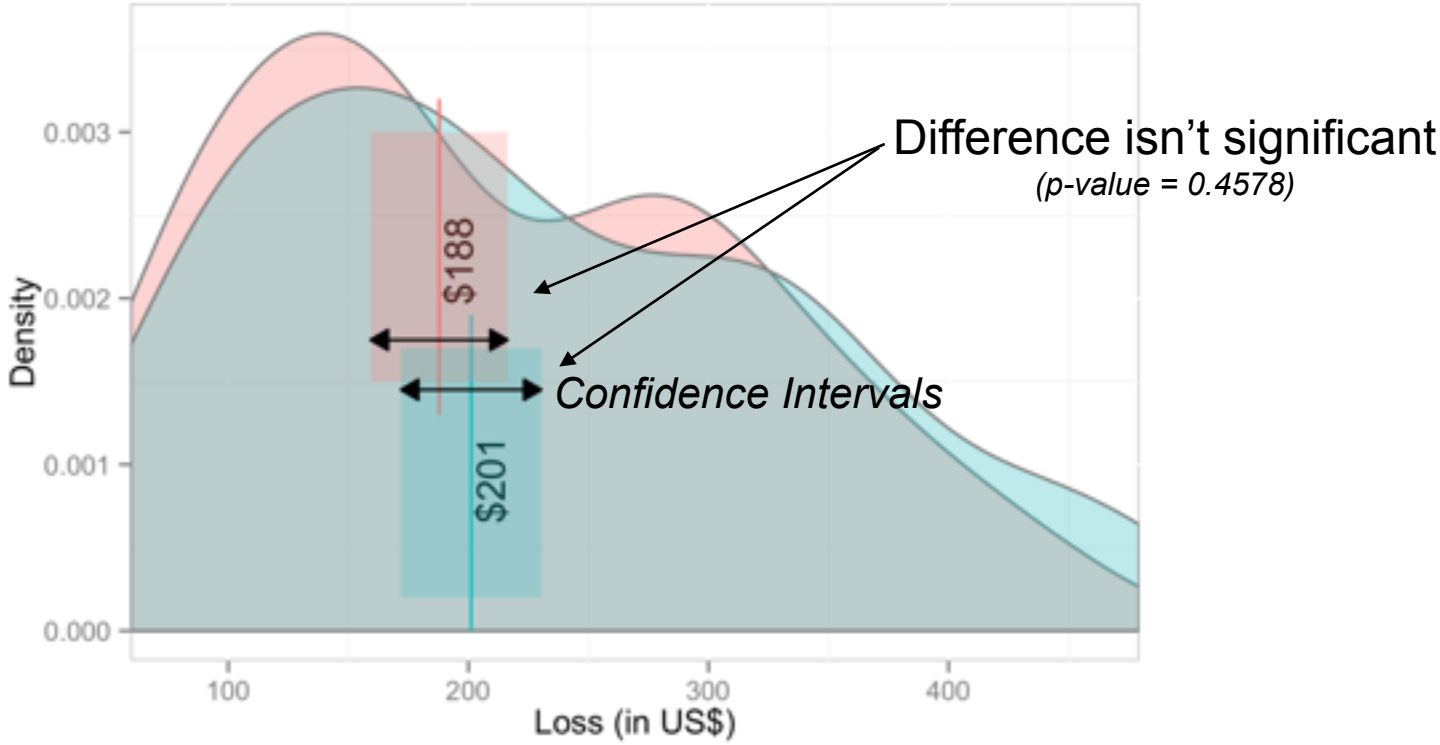
Ponemon finds increased abandonment by customers of organizations hit by data breaches



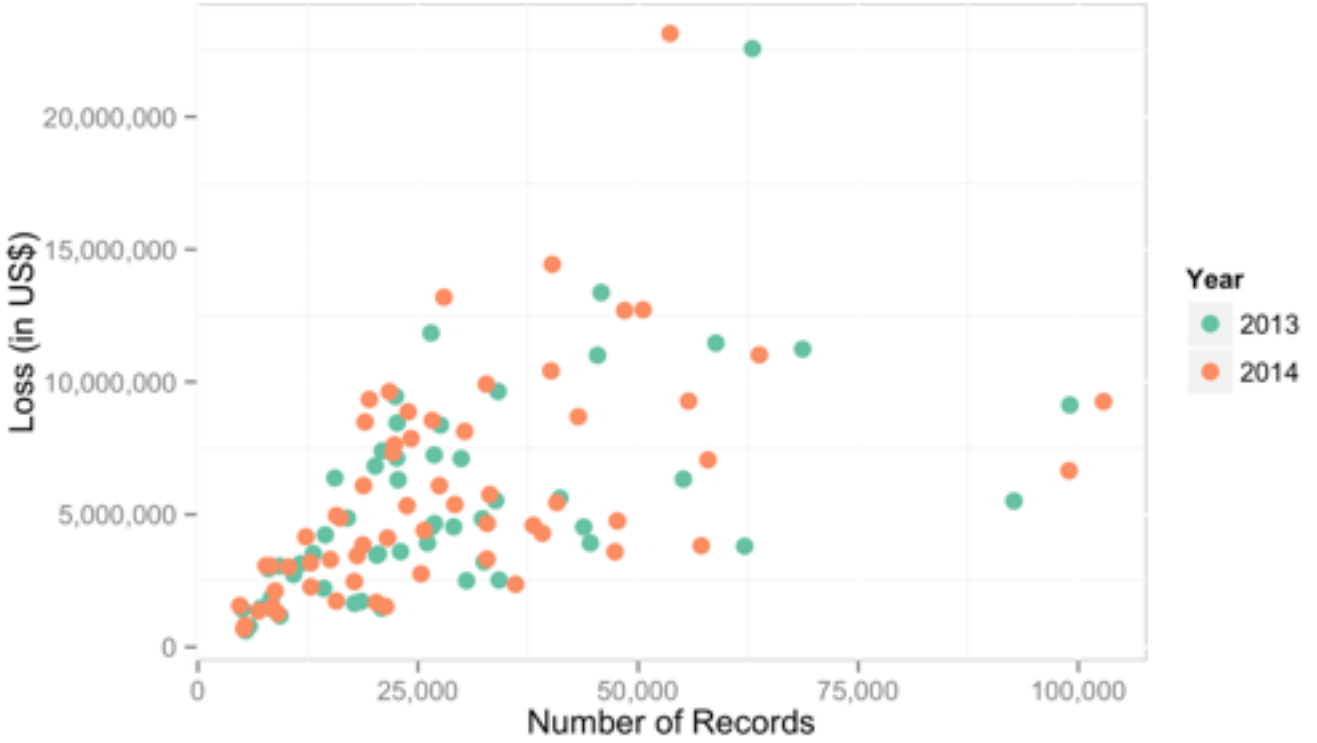
Least Squares to Ponemon



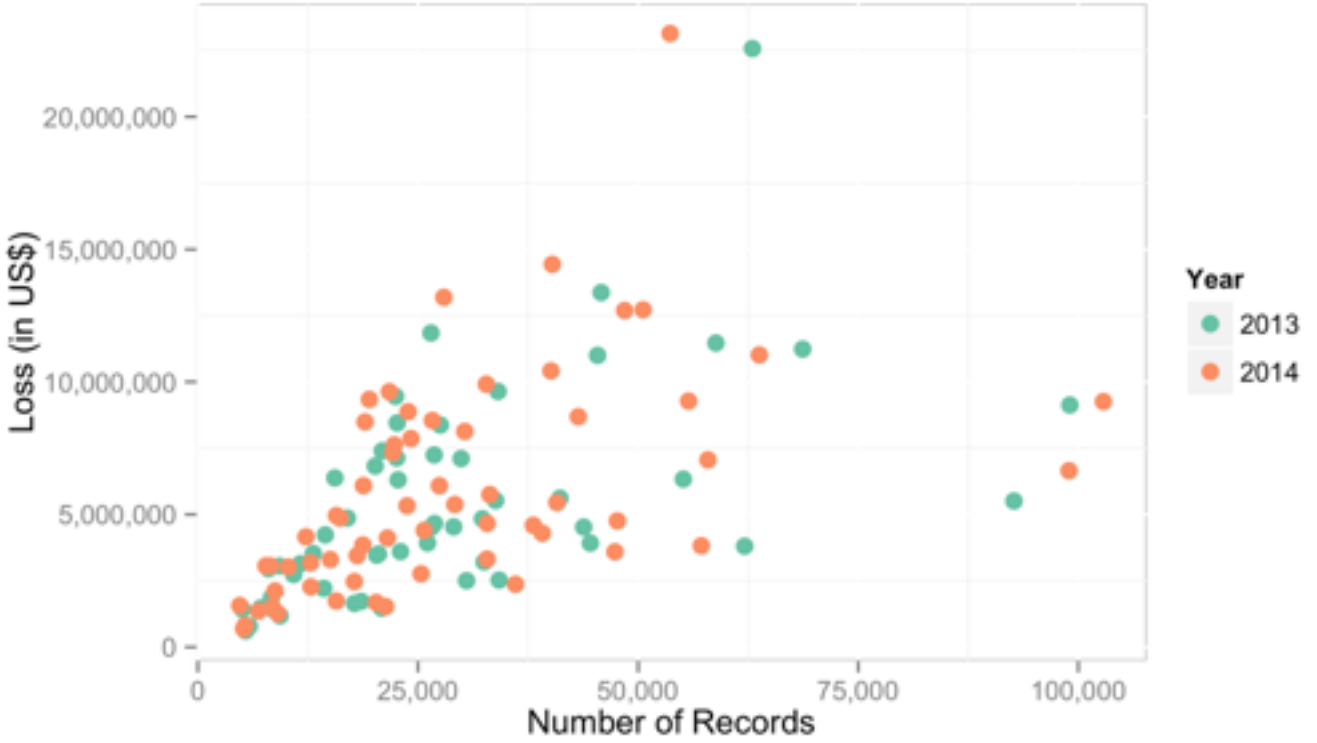
Least Squares to Ponemon



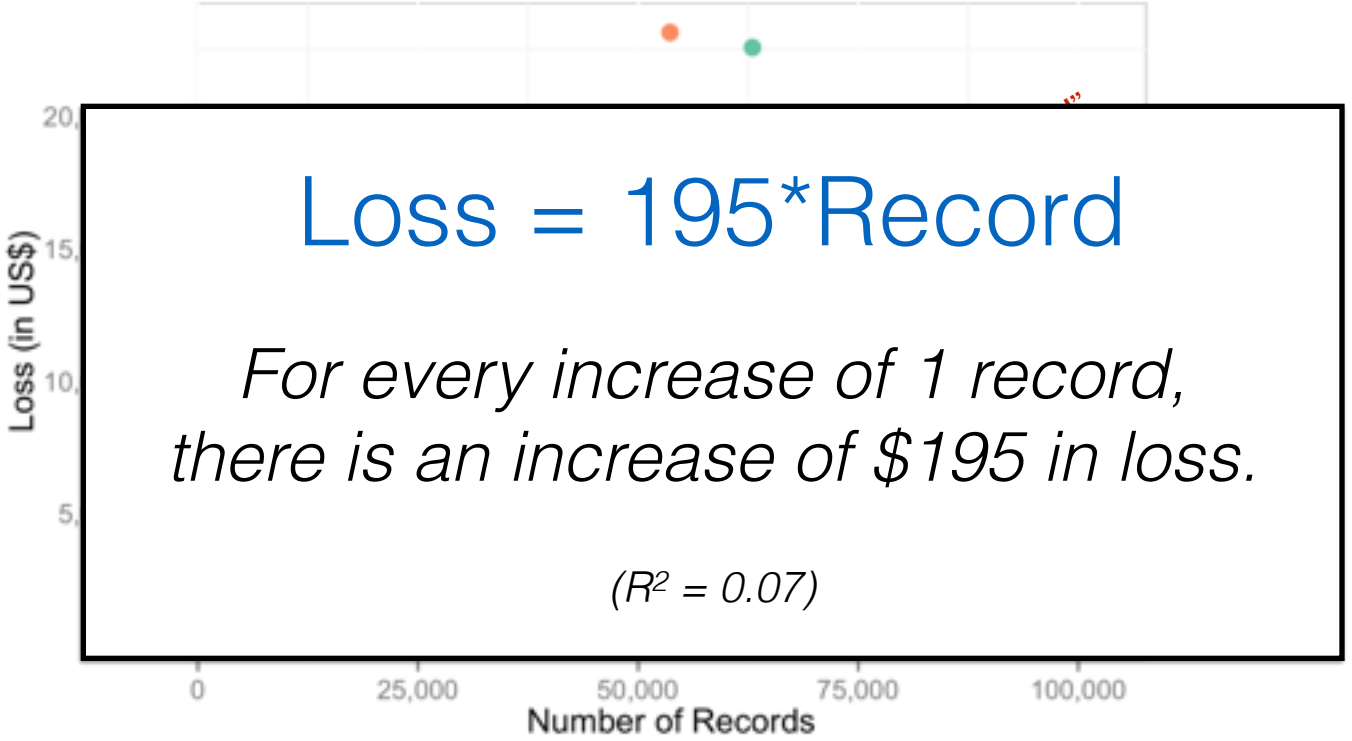
Least Squares to Ponemon



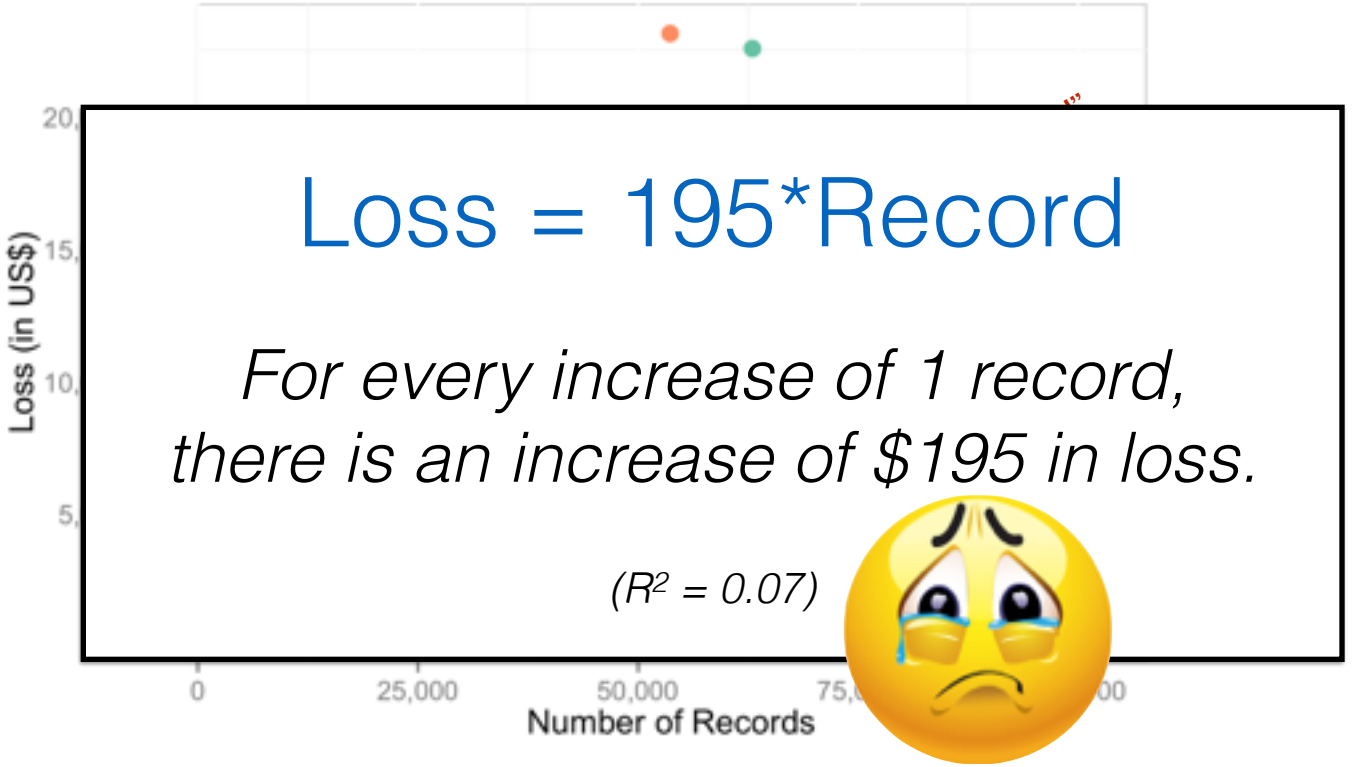
~~Least Squares to~~ Ponemon



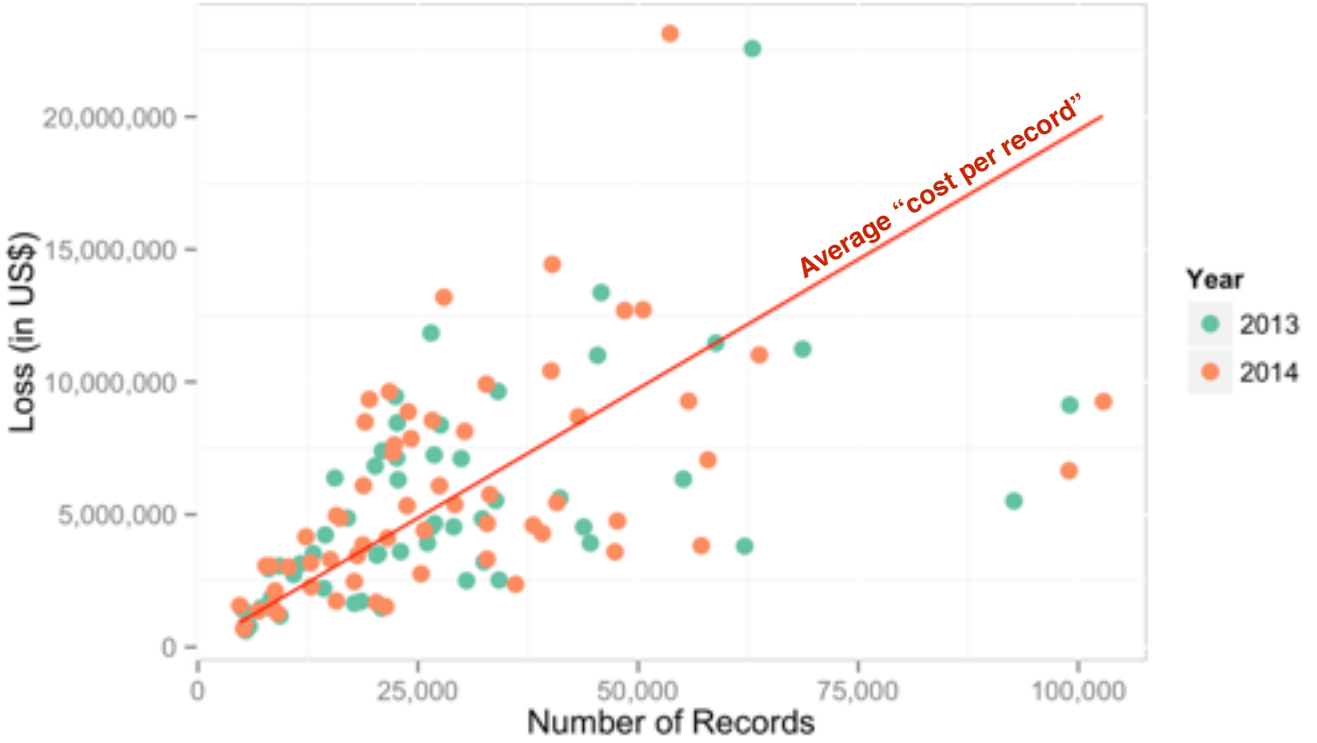
Least Squares to Ponemon



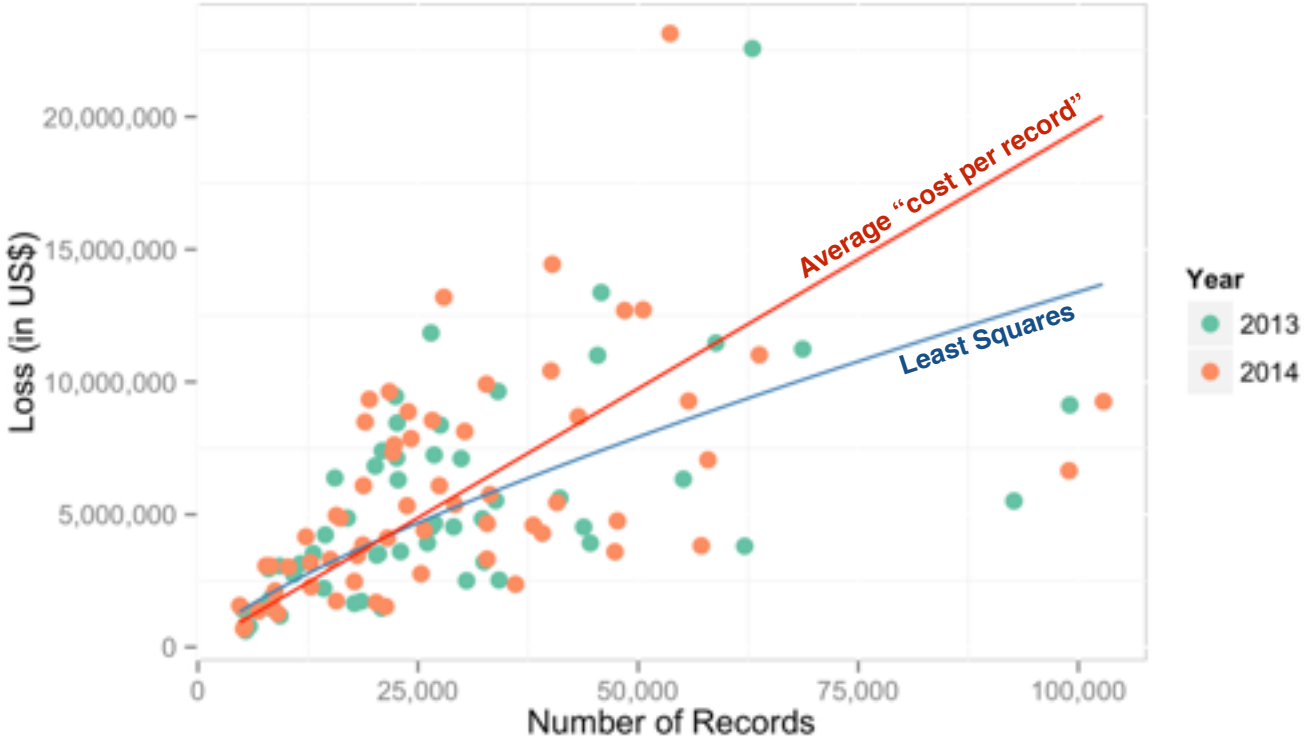
Least Squares to Ponemon



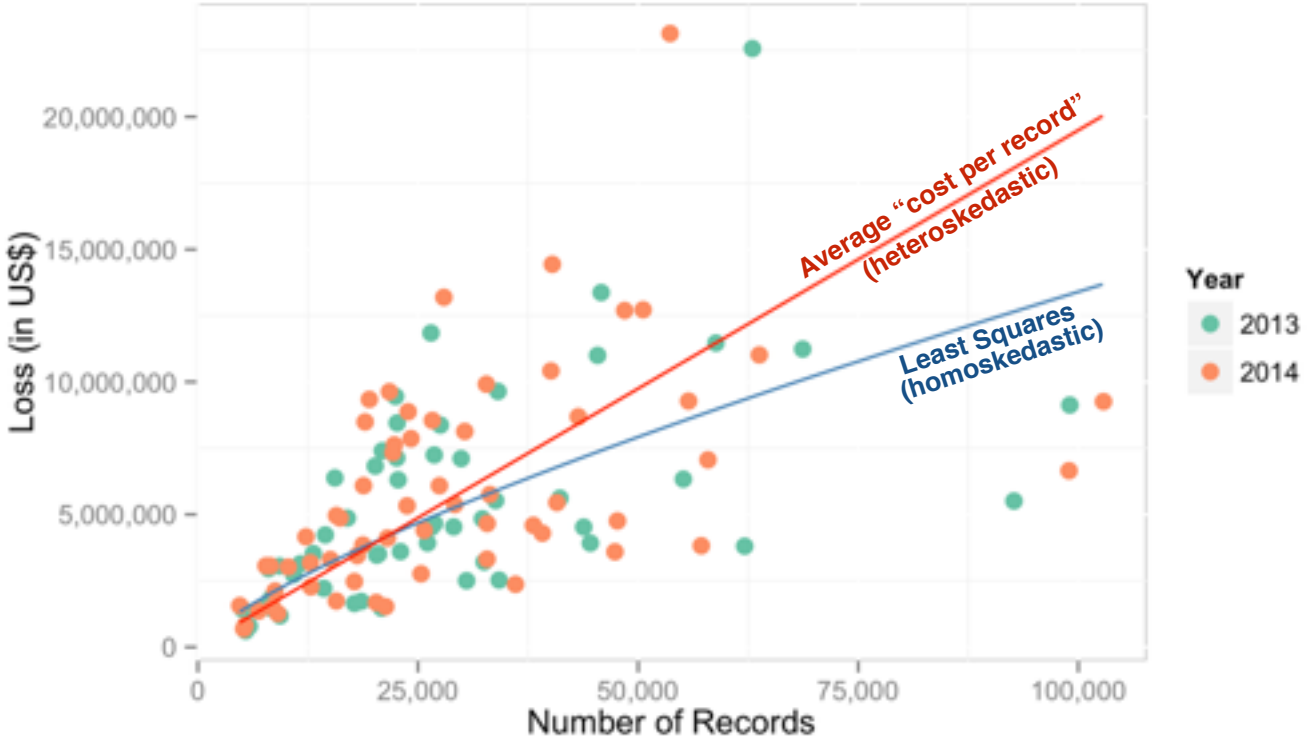
Least Squares to Ponemon



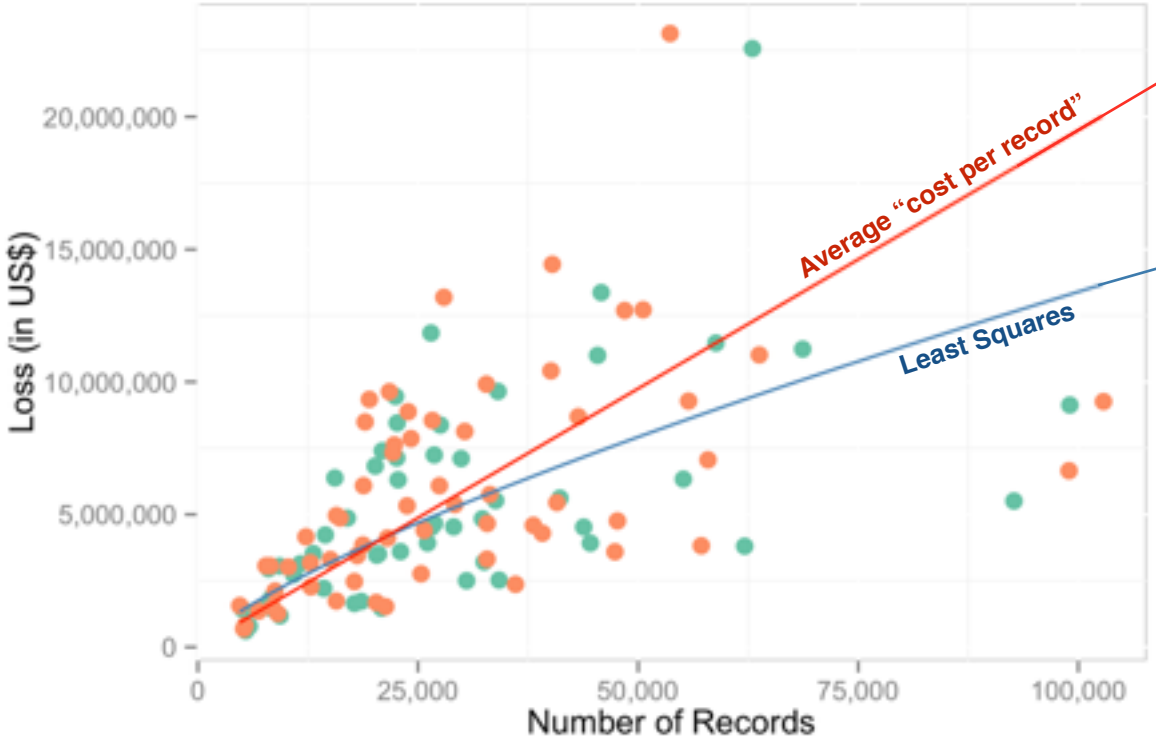
Least Squares to Ponemon



Least Squares to Ponemon



Least Squares to Ponemon



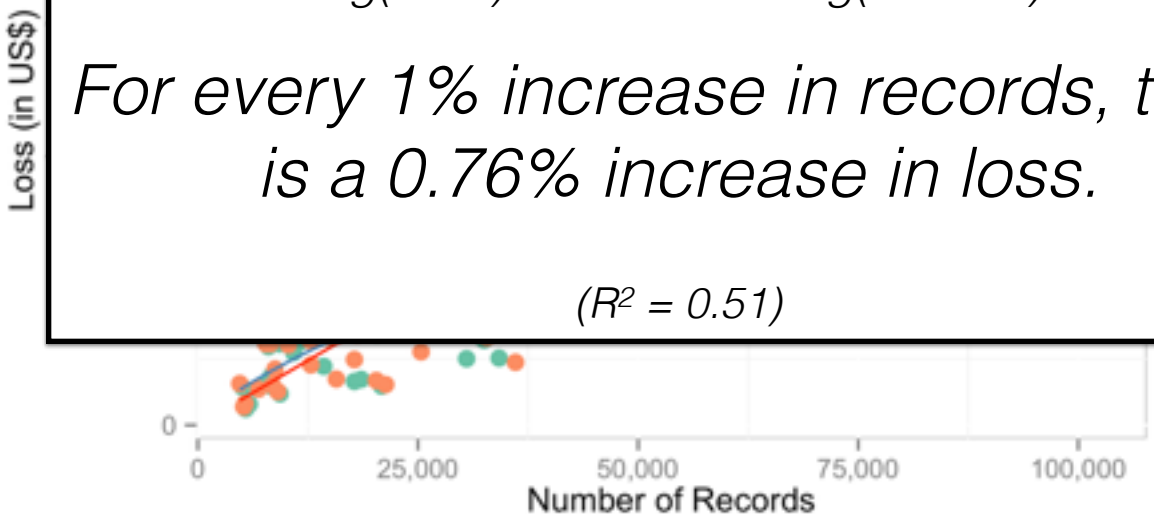
Least Squares to Ponemon

$$Loss = e^{(7.7 + 0.76*\log(Records))}$$

$$\log(Loss) = 7.7 + 0.76*\log(records)$$

For every 1% increase in records, there is a 0.76% increase in loss.

$(R^2 = 0.51)$



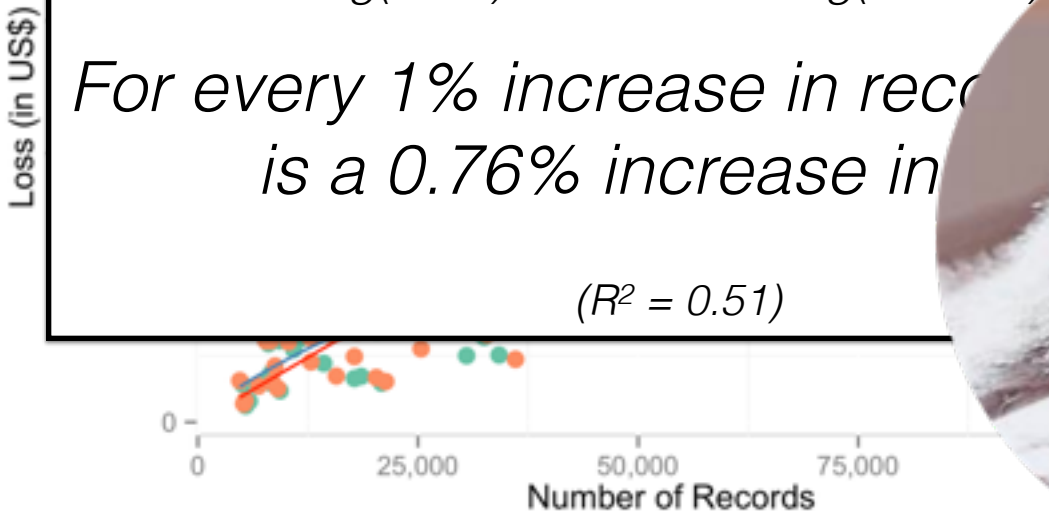
Least Squares to Ponemon

$$Loss = e^{(7.7 + 0.76*\log(Records))}$$

*log(Loss) = 7.7 + 0.76*log(records)*

For every 1% increase in reco
is a 0.76% increase in

(R^2 = 0.51)



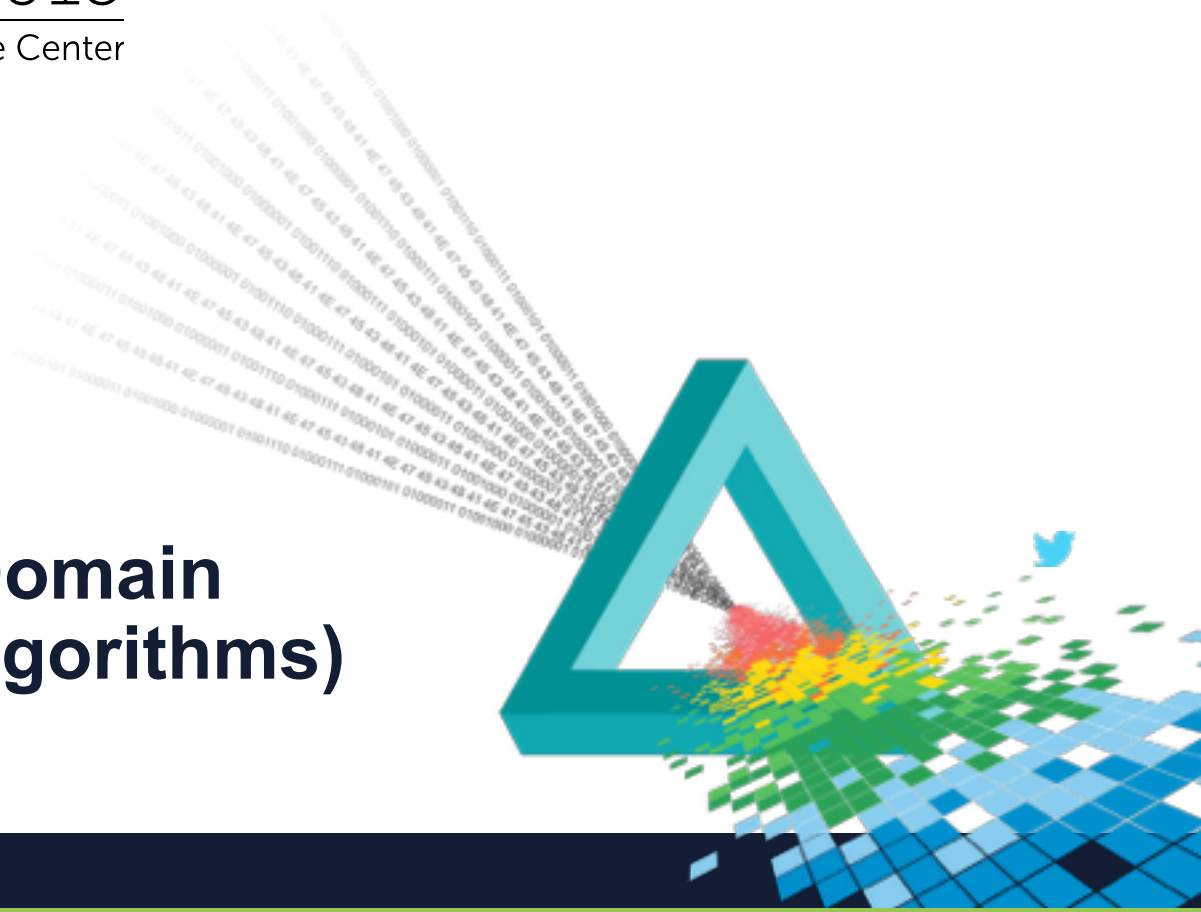
So What

- Regression (least squares) is the work horse of data analysis.
- Obvious and intuitive does not necessary mean it's right.
- Useful for quantitative variables… Collect data!

TRY THIS AT HOME!

- See DBIR for more detailed impact analysis
- See blog post for more Ponemon analysis – http://l.dds.ec/1CQHUa1

RSAConference2015
San Francisco | April 20-24 | Moscone Center



Talkin' 'bout my Domain GGGeneration (Algorithms)

Domain Generating Algorithms (DGA)

Algorithms that generate pseudo-random domain names. Used by malware to (typically) communicate with a controlling hub.

Cryptolocker	Goz	NewGoz	Legit
etledwndgunmrt	eiaupamojzhlrciwkeqhyxd	1erk1aq2tfv3e1dy8ikv1f0nxs8	fujifilm
obgfmoyfwptep	tkdabqnkrgdozhitdehypz	i5ep531lfuanc1ytynl1mmkio4	dallasdoglife
bugvesrwqxdjoa	uswodcmnvemqfmzxynjdnvhynvbe	zj7llmpk5fo87dtcg81e2j07c	startups
qxavdikemhepxk	ohhyhypphvgtucgiemfqdhai	vehvq1swdu9vuhfqvrcjxr46	askganesha
ohgnphscwbyvuse	ydqwmzhgaxoxfyzvcpvqgmfxro	1ncn8kn675d4o6dc4hh1f0se4r	wildcatdirectory
fbveqghechlth	kbcirszxzxscgeukcizjrntclvp	1v11tu8z5okt61njpiky1xoprmr	cherokeeherald
ihyrtyunnaltjm	eiseiondsgkbnzvgwdehxda	sd345o1rq011a1ms3qlley5yvu	admaster
auxiyeexsfcqj	ytwkpzlobljxkljhushyxkyt	1jz5ktklbpm53r2pdymmri043	directory2009
tknbivcmbekpwh	hswvovkduhlbfugqxpmfnjnzn	17adaod1oih6t91x358vyshspil	theupsstore
gtpjifumwmqpn	vwdjxoqworljhirgetwh	1e95km61jytx813ozodwofkggu	expediamail
cnqgglwrucrgp	xcbeeieymbguwddcabueipzwg	970z95v4nzg1qmt2c37ib43h	dyad-inc
aucdtwkdfyewc	pdqfrsvgkkfuwmvgpvvwayyzleu	5a3d2xgu8lq31bbf72q7l7o6c	qimaging

Classification…

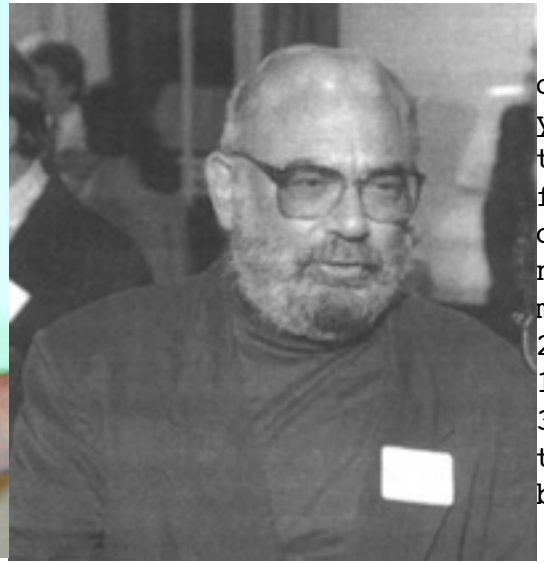
… is this one domain malicious?

Cryptolocker	Goz	NewGoz	Legit
etledwndgunmrt	eiaupamojzhlrciwkeqhyxd	1erk1aq2tfv3e1dy8ikv1f0nxs8	fujifilm
obgfmoyfwptep	tkdabqnkrgdozhitdehypz	i5ep531lfuanc1ytynl1mmkio4	dallasdoglife
bugvesrwqxdjoa	uswodcmnvemqfmzxynjdnvhynvbe	zj7llmpk5fo87dtcg81e2j07c	startups
qxavdikemhepxk	ohhyhypphvgtucgiemfqdhai	vehvq1swdu9vuhfqvrcjxr46	askganesha
ohgnphscwbyvuse	ydqwmzhgaxoxfyzvcpvqgmfxro	1ncn8kn675d4o6dc4hh1f0se4r	wildcatdirectory
fbveqghechlth	kbcirszxzxscgeukcizjrntclvp	1v11tu8z5okt61njpiky1xoprmr	cherokeeherald
ihyrtyunnaltjm	eiseiondsgkbnzvgwdehxda	sd345o1rq011a1ms3qlley5yvu	admaster
auxiyeexsfcqj	ytwkpzlobljxkljhushyxkyt	1jz5ktklbpm53r2pdymmri043	directory2009
tknbivcmbekpwh	hswvovkduhlbfugqxpmfnjnzn	17adaod1oih6t91x358vyshspil	theupsstore
gtpjifumwmqpn	vwdjxoqworljhirgetwh	1e95km61jytx813ozodwofkggu	expediamail
cnqgglwrucrgp	xcbeeieymbguwddcabueipzwg	970z95v4nzg1qmt2c37ib43h	dyad-inc
aucdtwkdfyewc	pdqfrsvgkkfuwmvgpvvwayyzleu	5a3d2xgu8lq31bbf72q7l7o6c	qimaging

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman
1928-2005



C	Legit
et	dy8ikv1f0nxs8	fujifilm
ob	ytynl1mmkio4	dallasdoglife
bu	tcg81e2j07c	startups
qz	fqvrcjxr46	askganesha
ob	dc4hh1f0se4r	wildcatdirectory
fb	njpiky1xoprmr	cherokeeherald
ib	ms3qlley5yvu	admaster
au	2pdymmri043	directory2009
tk	1x358vyshspil	theupsstore
gt	3ozodwofkggu	expediamail
cr	t2c37ib43h	dyad-inc
au	bf72q7l7o6c	qimaging

Features (machine learning)

Cryptolocker	Goz	NewGoz	Legit
etledwndgunmrt	eiaupamojzhlrciwkeqhyxd	1erk1aq2tfv3e1dy8ikv1f0nxs8	fujifilm
obgfmoyfwptep	tkdabqnkrgdozhitdehypz	i5ep531lfuanc1ytynl1mmkio4	dallasdoglife
bugvesrwqxdjoa	uswodcmnvemqfmzxynjdnvhynvbe	zj7llmpk5fo87dtcg81e2j07c	startups
qxavdikemhepxk	ohhyhypphvgtucgiemfqdhai	vehvq1swdu9vuhfqvrcjxr46	askganesha
ohgnphscwbyvuse	ydqwmzhgaxoxfyzvcpvqgmfxro	1ncn8kn675d4o6dc4hh1f0se4r	wildcatdirectory
fbveqghechlth	kbcirszxzxscgeukcizjrntclvp	1v11tu8z5okt61njpiky1xoprmr	cherokeeherald
ihyrtyunnaltjm	eiseiondsgkbnzvgwdehxda	sd345o1rq011a1ms3qlley5yvu	admaster
auxiyeexsfcqj	ytwkpzlobljxkljhushyxkyt	1jz5ktklbpm53r2pdymmri043	directory2009
tknbivcmbekpwh	hswvovkduhlbfugqxpmfnjnzn	17adaod1oih6t91x358vyshspil	theupsstore
gtpjifumwmqpn	vwdjxoqworljhirgetwh	1e95km61jytx813ozodwofkggu	expediamail
cnqgglwrucrgp	xcbeeieymbguwddcabueipzwg	970z95v4nzg1qmt2c37ib43h	dyad-inc
aucdtwkdfyewc	pdqfrsvgkkfuwmvgpvvwayyzleu	5a3d2xgu8lq31bbf72q7l7o6c	qimaging

Features (machine learning)

- Length
- Entropy
- letter sequences (n-grams)
- Others?

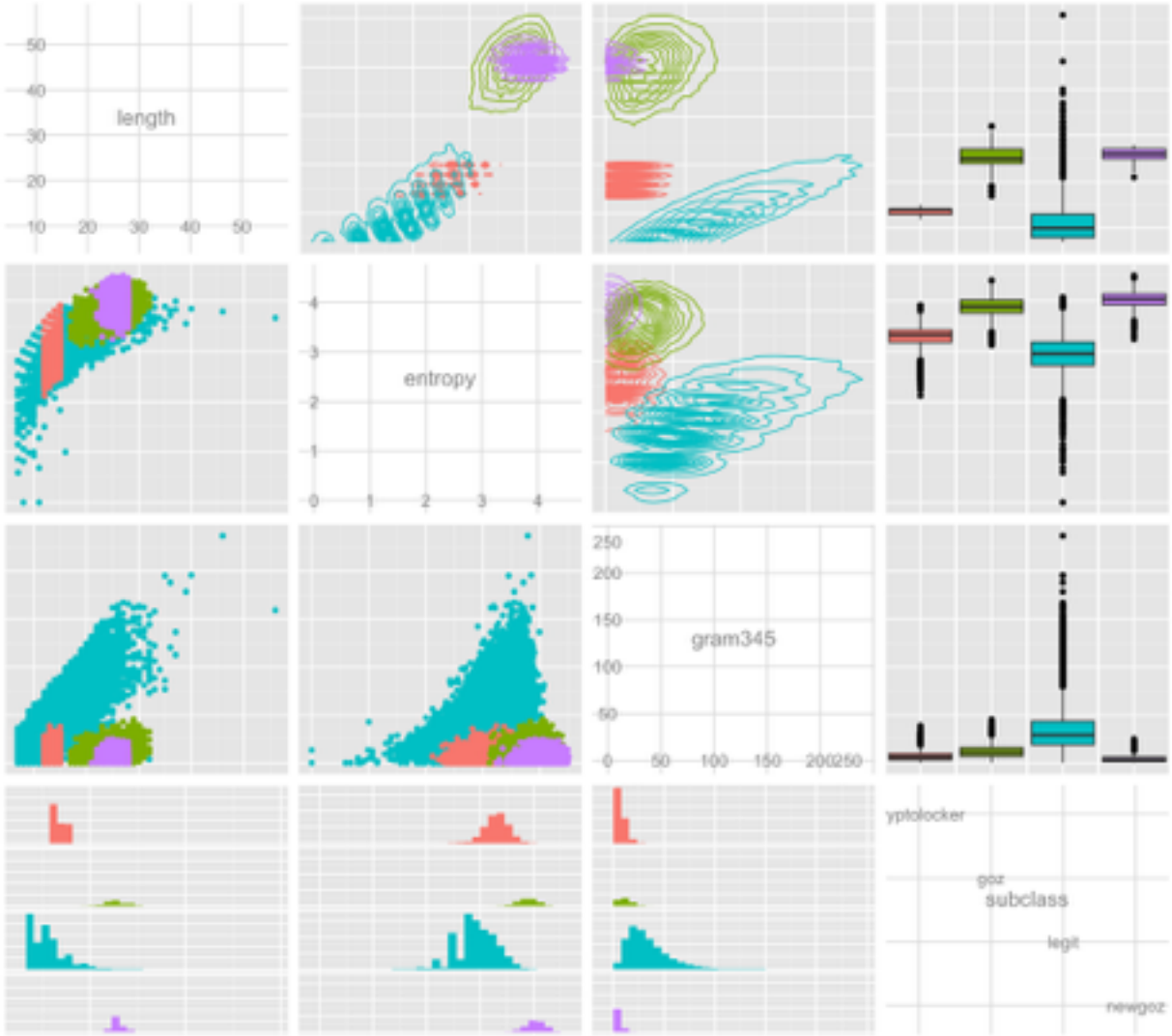
Cryptolocker	Goz	NewGoz	Legit
etledwndgunmrt	eiaupamojzhlrciwkeqhyxd	1erk1aq2tfv3e1dy8ikv1f0nxs8	fujifilm
obgfmoyfwptep	tkdabqnkrgdozhitdehypz	i5ep531lfuanc1ytynl1mmkio4	dallasdoglife
bugvesrwqxdjoa	uswodcmnvemqfmzxynjdnvhynvbe	zj7llmpk5fo87dtcg81e2j07c	startups
qxavdikemhepxk	ohhyhypphvgtucgiemfqdhai	vehvq1swdu9vuhfqvrcjxr46	askganesha
ohgnphscwbyvuse	ydqwmzhgaxoxfyzvcpvqgmfxro	1ncn8kn675d4o6dc4hh1f0se4r	wildcatdirectory
fbveqghechlth	kbcirszxzxscgeukcizjrntclvp	1v11tu8z5okt61njpiky1xoprmr	cherokeeherald
ihyrtyunnaltjm	eiseiondsgkbnzvgwdehxda	sd345o1rq011a1ms3qlley5yvu	admaster
auxiyeexsfcqj	ytwkpzlobljxkljhushyxkyt	1jz5ktklbpm53r2pdymmri043	directory2009
tknbivcmbekpwh	hswvovkduhlbfugqxpmfnjnzn	17adaod1oih6t91x358vyshspil	theupsstore
gtpjifumwmqpn	vwdjxoqworljhirgetwh	1e95km61jytx813ozodwofkggu	expediamail
cnqgglwrucrgp	xcbeeieymbguwddcabueipzwg	970z95v4nzg1qmt2c37ib43h	dyad-inc
aucdtwkdfyewc	pdqfrsvgkkfuwmvgpvvwayyzleu	5a3d2xgu8lq31bbf72q7l7o6c	qimaging

The Features (what they look like)

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

Comparing all the Features…



The Results

	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

	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

The Results (in the gray area)

	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

```
          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
```

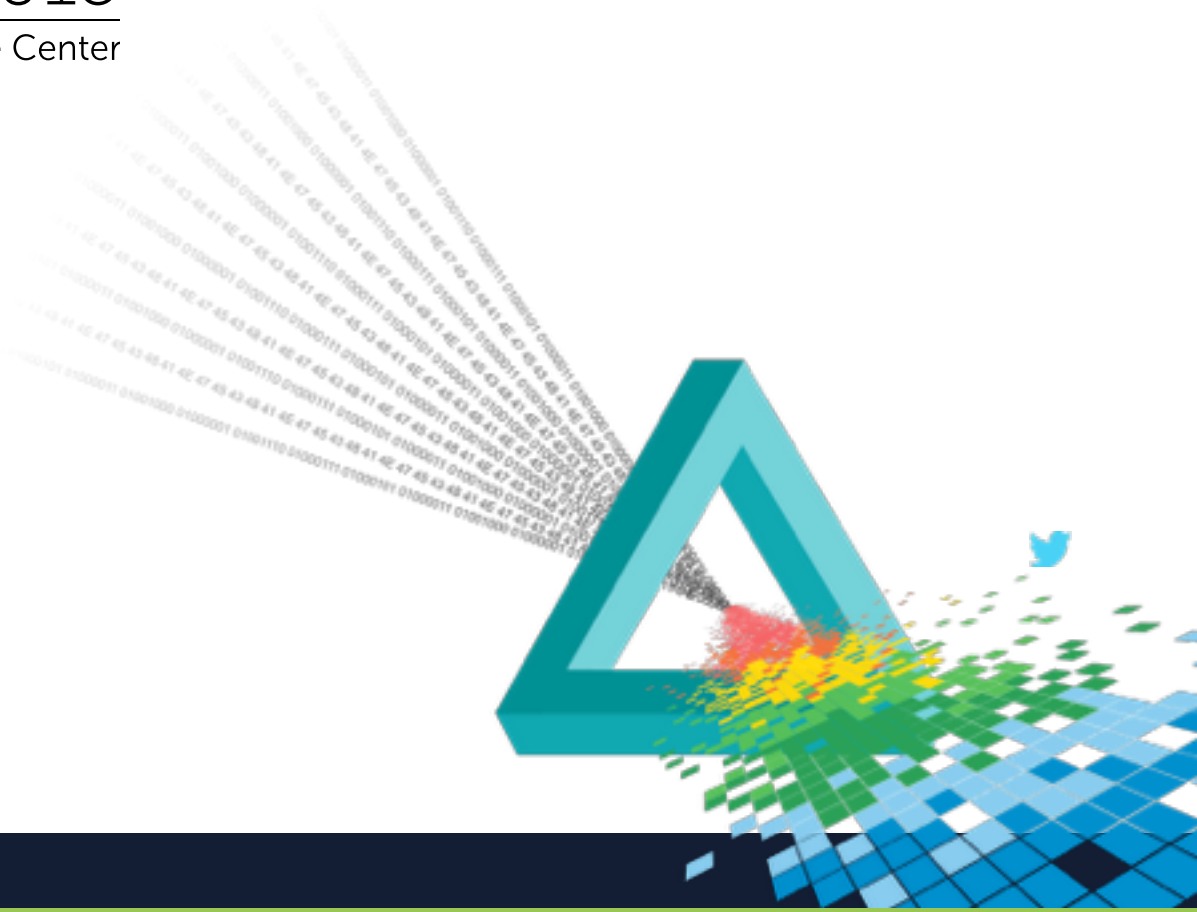
```
            Sensitivity : 0.9948
            Specificity : 0.9956
         Pos Pred Value : 0.9929
         Neg Pred Value : 0.9968
             Prevalence : 0.3789
         Detection Rate : 0.3769
   Detection Prevalence : 0.3797
      Balanced Accuracy : 0.9952
```

So What

- DNS is a rich source of data in your enterprise (and it's FREE)
- Can collect it from logs, sniffed off white, even retrieved from latest netflow standard
- Can potentially give you a leg up on targeted attacks specific to only your org

- See blog post(s) for more DGA analysis

You have permission to do this



Questions?