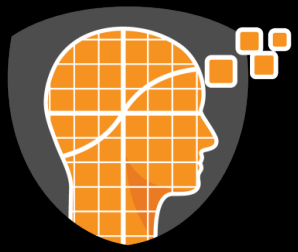


MLSec
MACHINE LEARNING SECURITY

NIDDEL™

Data-Driven Threat Intelligence: Useful Methods and Measurements for Handling Indicators (#ddti)

Alex Pinto
Chief Data Scientist
Niddel / MLSec Project
@alexcpsec
@MLSecProject

Alexandre Sieira
CTO
Niddel
@AlexandreSieira
@NiddelCorp

MLSec Project / Niddel

- MLSec Project – research-focused branch of Niddel for open-source tools and community building
- Niddel builds Magnet, an Applied Threat Intelligence Platform, that leverages ML models to facilitate the usage of TI in a fully personalized and hassle-free way.
- Looking for beta testers and research collaboration
- More info at:
  - **niddel.com**
  - **mlsecproject.org**

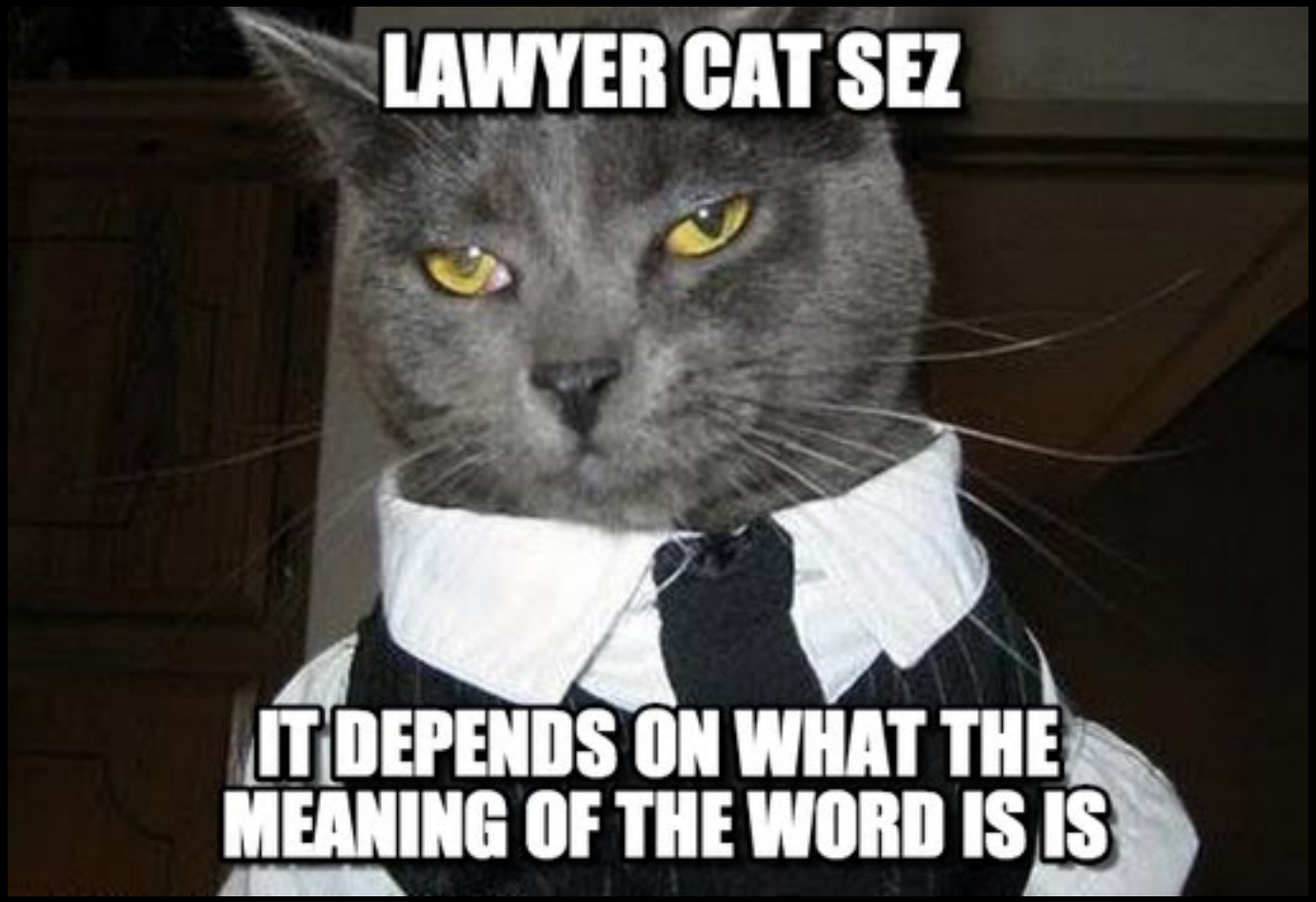
MLSec
MACHINE LEARNING SECURITY

NIDDEL™

Agenda

- ~~Cyber War…~~ Threat Intel – What is it good for?
- Combine and TIQ-test
- Using TIQ-test
  - Novelty Test
  - Overlap Test
  - Population Test
  - Aging Test
  - Uniqueness Test
- Use case: Feed Comparison

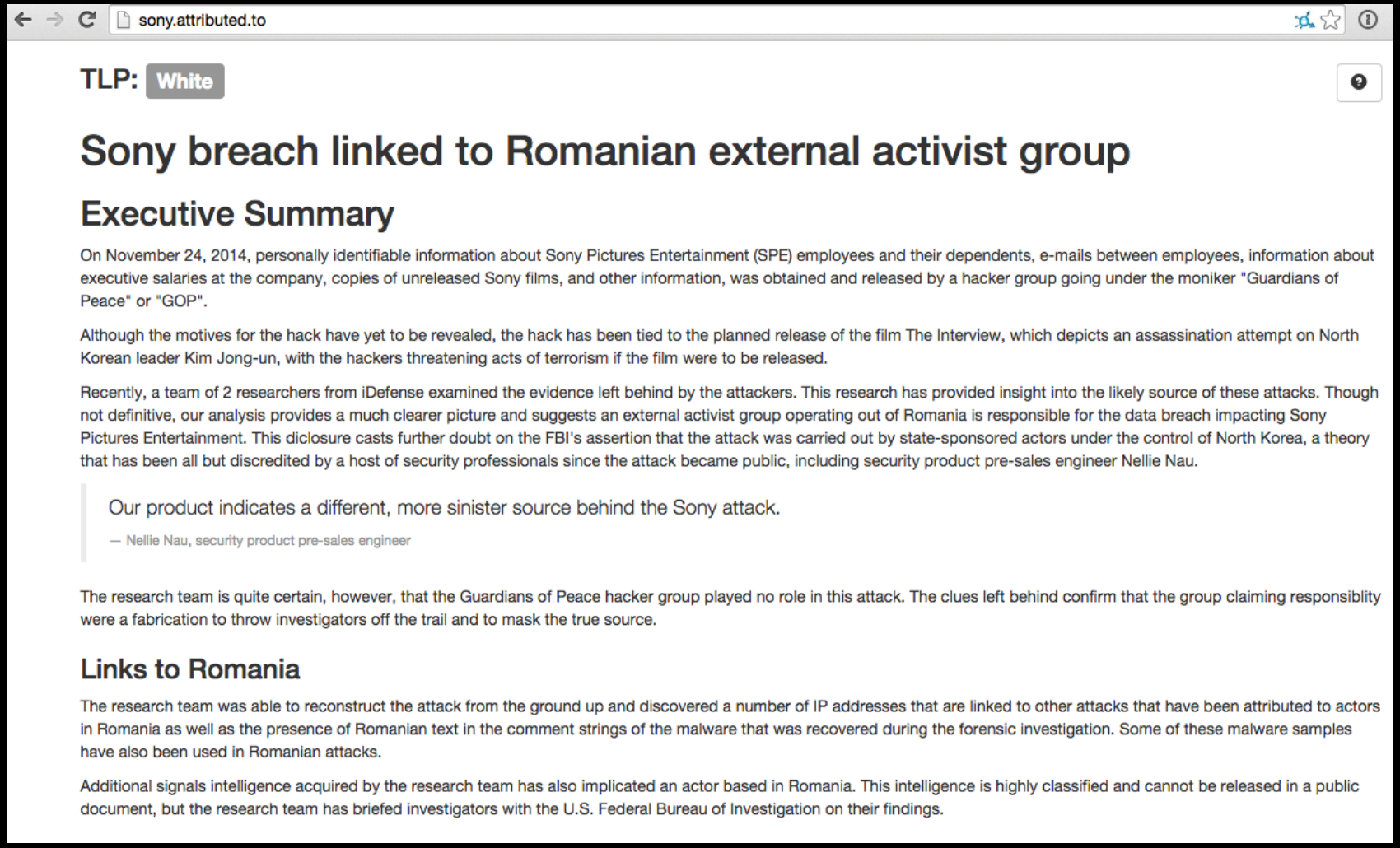
What is TI good for anyway?



What is TI good for anyway?

- 1) Attribution

What is TI good for anyway?



What is TI good for anyway?

- 2) Cyber Threat Maps

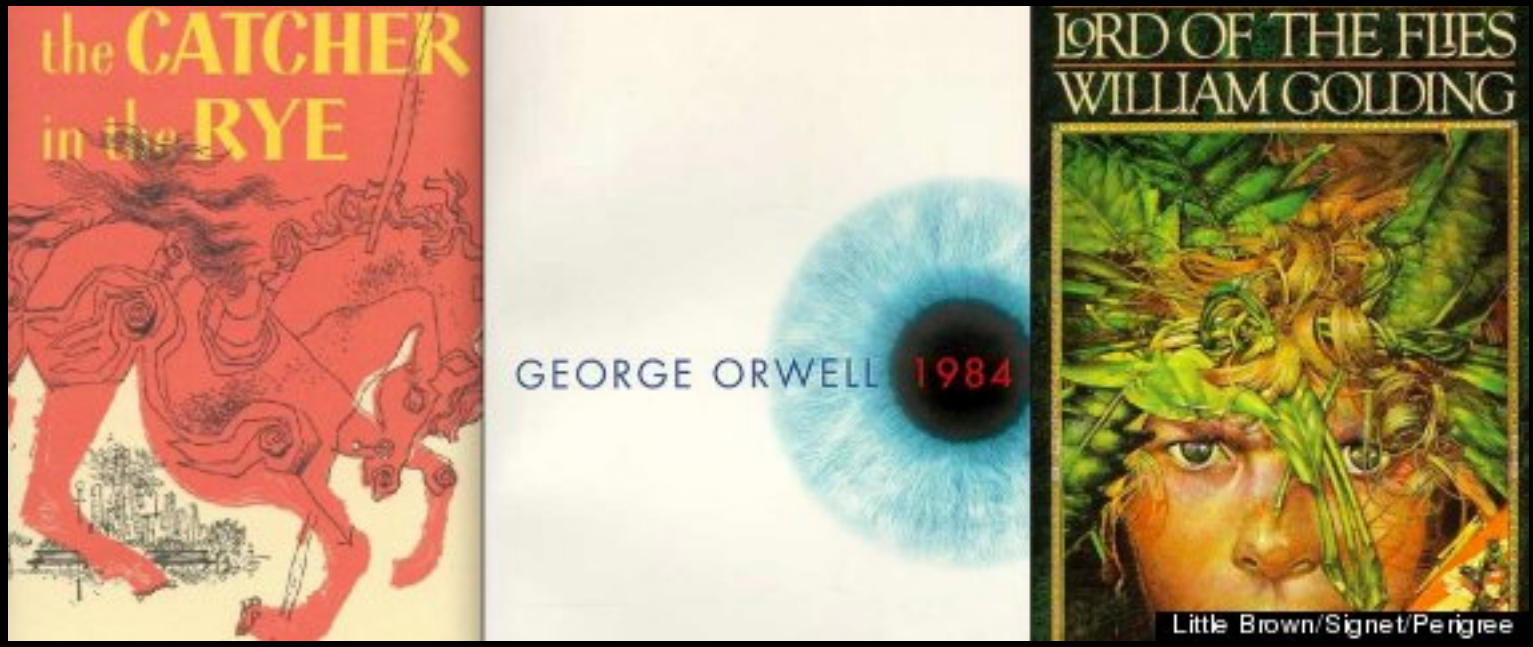
  (https://github.com/hrbrmstr/pewpew)

What is TI good for anyway?

- 3) How about actual defense?
  - Use it as blacklists? As research data?
  - Thing is RAW DATA is hard to work with

(Semi-)Required Reading

- #tiqtest slides: http://bit.ly/tiqtest
- RPubs Page: http://bit.ly/tiqtest-rpubs



Combine and TIQ-Test

- Combine (https://github.com/mlsecproject/combine)
  - Gathers TI data (ip/host) from Internet and local files
  - Normalizes the data and enriches it (AS / Geo / pDNS)
  - Can export to CSV, "tiq-test format" and CRITs
  - Coming soon: CybOX / STIX (ty @kylemaxwell)

- TIQ-Test (https://github.com/mlsecproject/tiq-test)
  - Runs statistical summaries and tests on TI feeds
  - Generates charts based on the tests and summaries
  - Written in R (because you should learn a stat language)

Using TIQ-TEST

- Available tests and statistics:
  - NOVELTY – How often do they update themselves?
  - OVERLAP – How do they compare to what you got?
  - POPULATION – How does this population distribution compare to another one ?
  - AGING – How long does an indicator sit on a feed?
  - UNIQUENESS – How many indicators are found in only one feed?

Using TIQ-TEST

- New dataset!
- https://github.com/mlsecproject/tiq-test-Winter2015

```
print(tiq.data.getAvailableDates("raw", "public_outbound"))
```

```
##  [1] "20141001" "20141002" "20141003" "20141004" "20141005" "20141006"
##  [7] "20141007" "20141008" "20141009" "20141010" "20141011" "20141012"
## [13] "20141013" "20141014" "20141015" "20141016" "20141017" "20141018"
## [19] "20141019" "20141020" "20141021" "20141022" "20141023" "20141024"
## [25] "20141025" "20141026" "20141027" "20141028" "20141029" "20141030"
## [31] "20141031" "20141101" "20141102" "20141103" "20141104" "20141105"
## [37] "20141106" "20141107" "20141108" "20141109" "20141110" "20141111"
## [43] "20141112" "20141113" "20141114" "20141115" "20141116" "20141117"
## [49] "20141118" "20141119" "20141120" "20141121" "20141122" "20141123"
## [55] "20141124" "20141125" "20141126" "20141127" "20141128" "20141129"
## [61] "20141130"
```

Using TIQ-TEST – Feeds Selected

- Dataset was separated into “inbound” and “outbound”

```
inbound.ti = tiq.data.loadTI("raw", "public_inbound", "20141101")
unique(inbound.ti$source)
```

```
##  [1] "alienvault"        "autoshun"          "blocklistde"
##  [4] "botscout"          "bruteforceblocker" "charleshaley"
##  [7] "ciarmy"            "dragonresearch"    "dshield"
## [10] "honeypot"          "openbl"            "packetmail"
## [13] "virbl"
```

```
outbound.ti = tiq.data.loadTI("raw", "public_outbound", "20141101")
unique(outbound.ti$source)
```

```
##  [1] "alienvault"        "feodo"             "malcode"
##  [4] "malcode_zones"     "malwaredomainlist" "malwaredomains"
##  [7] "malwaregroup"      "palevotracker"     "spyeye"
## [10] "sslbl"             "zeus"
```

Using TIQ-TEST – Data Prep

- Extract the “raw” information from indicator feeds
- Both IP addresses and hostnames were extracted

```
outbound.ti = tiq.data.loadTI("raw", "public_outbound", "20141101")
outbound.ti[, list(entity, type, direction, source, date)]
```

```
##                          entity type direction     source       date
##     1:             1.168.15.140 IPv4  outbound alienvault 2014-11-01
##     2:                1.93.6.86 IPv4  outbound alienvault 2014-11-01
##     3:             100.42.211.4 IPv4  outbound alienvault 2014-11-01
##     4:           101.227.172.24 IPv4  outbound alienvault 2014-11-01
##     5:             101.36.81.55 IPv4  outbound alienvault 2014-11-01
##    ---
## 11388:          up.frigo2000.it FQDN  outbound       zeus 2014-11-01
## 11389:          update.odeen.eu FQDN  outbound       zeus 2014-11-01
## 11390: update.rifugiopontese.it FQDN  outbound       zeus 2014-11-01
## 11391:       vahendkarasis4.com FQDN  outbound       zeus 2014-11-01
## 11392:           welcahllyn.com FQDN  outbound       zeus 2014-11-01
```

Using TIQ-TEST – Data Prep

- Convert the hostname data to IP addresses:
  - Active IP addresses for the respective date ("A" query)
  - Passive DNS from Farsight Security (DNSDB)

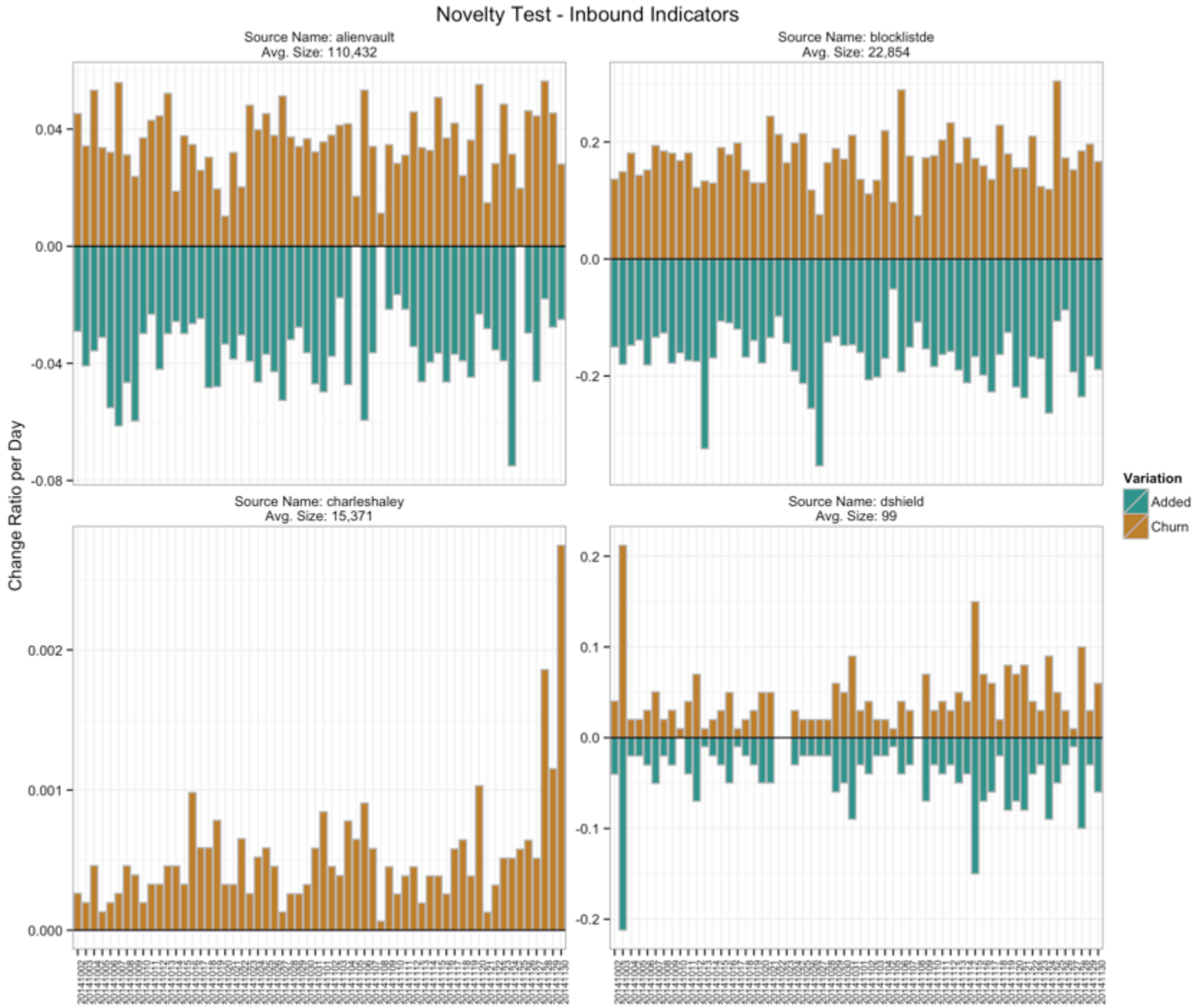
- For each IP record (including the ones from hostnames):
  - Add asnumber and asname (from MaxMind ASN DB)
  - Add country (from MaxMind GeoLite DB)
  - Add rhost (again from DNSDB) – most popular "PTR"

Using TIQ-TEST – Data Prep Done

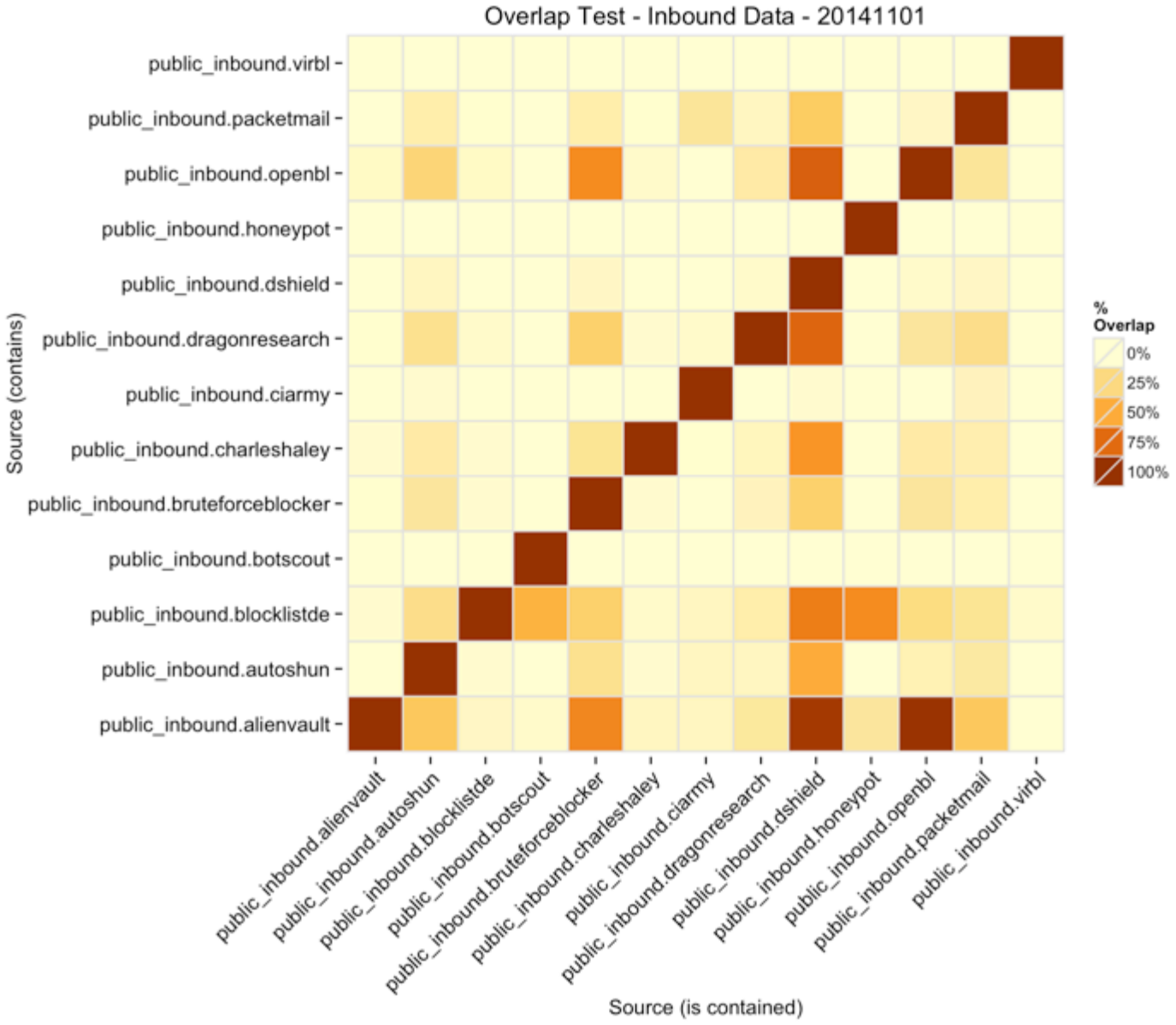
```
enrich.ti = tiq.data.loadTI("enriched", "public_outbound", "20141101")
enrich.ti = enrich.ti[, notes := NULL]
tail(enrich.ti)
```

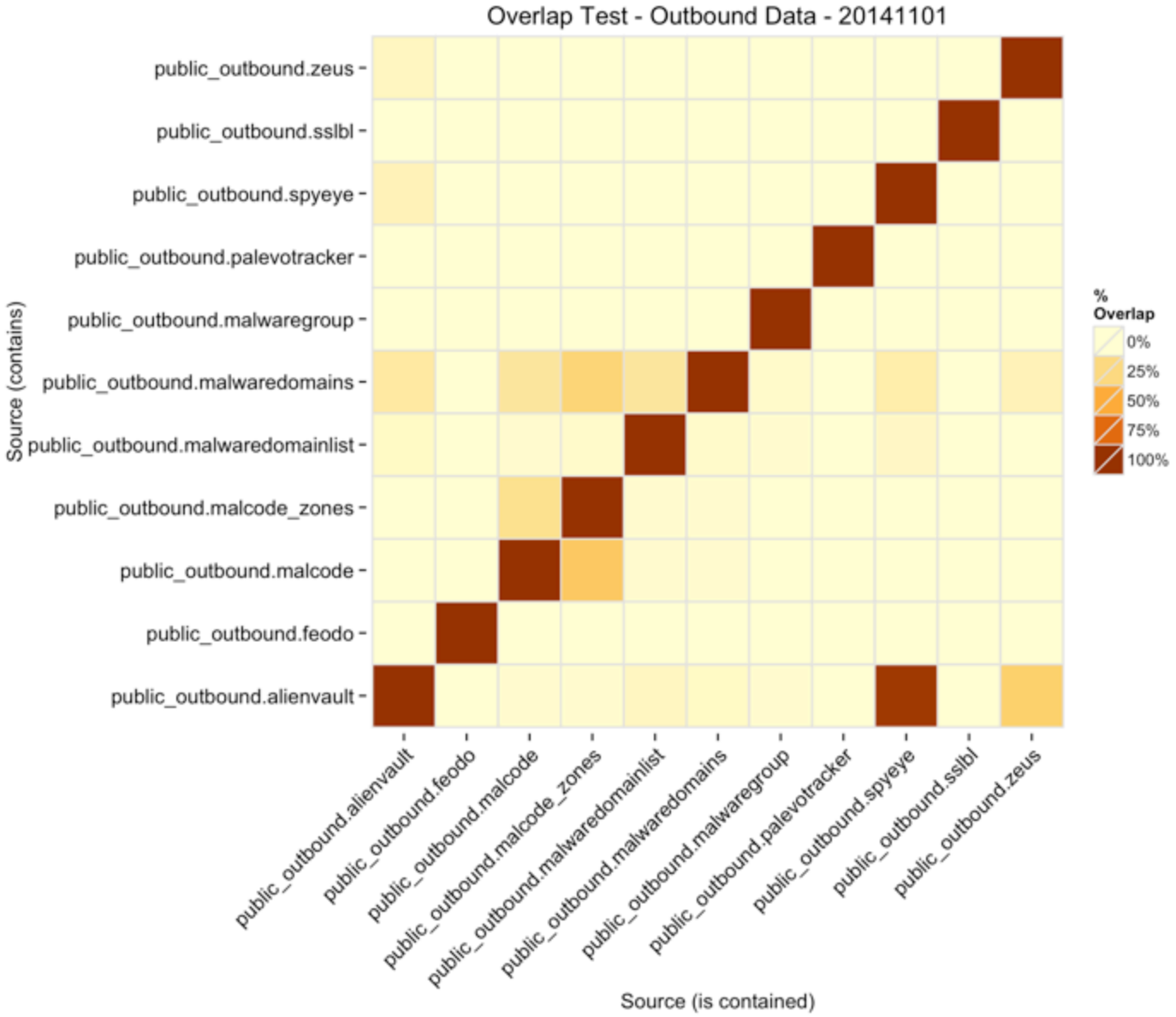
```
##            entity type direction source       date asnumber
## 1:  94.102.63.153 IPv4  outbound   zeus 2014-11-01    29073
## 2:   94.103.36.55 IPv4  outbound   zeus 2014-11-01    47894
## 3:  95.163.121.12 IPv4  outbound   zeus 2014-11-01    12695
## 4: 98.131.185.136 IPv4  outbound   zeus 2014-11-01    32392
## 5: 98.131.185.136 IPv4  outbound   zeus 2014-11-01    32392
## 6:    99.181.5.83 IPv4  outbound   zeus 2014-11-01     7018
##                     asname country                         host
## 1:          Ecatel Network      NL                           NA
## 2: VeriTeknik Bilisim Ltd.      TR                           NA
## 3:   Digital Networks CJSC      RU                           NA
## 4:   Ecommerce Corporation      US                           NA
## 5:   Ecommerce Corporation      US projects.globaltronics.net
## 6:     AT&T Services, Inc.      US                           NA
##                                      rhost
## 1:                           exadomains.net
## 2:                datacenter.veriteknik.com
## 3:                                     NA
## 4:                                     NA
## 5:                                     NA
## 6: adsl-99-181-5-83.dsl.irvnca.sbcglobal.net
```

Novelty Test – measuring added and dropped indicators


Novelty Test - Inbound Indicators
Source Name: alienvault
Avg. Size: 110,432
Source Name: blocklistde
Avg. Size: 22,854
Source Name: charleshaley
Avg. Size: 15,371
Source Name: dshield
Avg. Size: 99
Change Ratio per Day
0.04
0.00
-0.04
-0.08
0.2
0.0
-0.2
0.002
0.001
0.000
0.1
-0.1
Variation
Added
Churn

Overlap Test – More data is better, but make sure it is not the same data

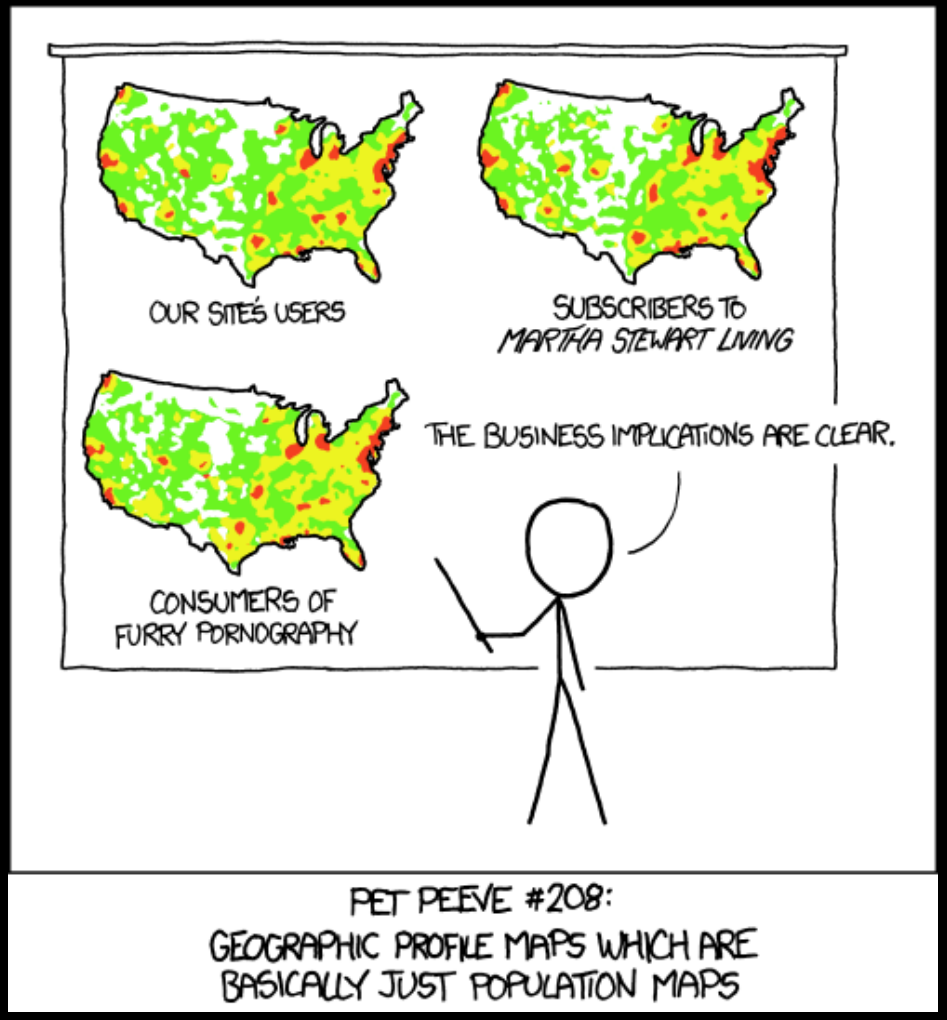

Overlap Test - Inbound Data - 20141101
Source (contains)
public_inbound.virbl
public_inbound.packetmail
public_inbound.openbl
public_inbound.honeypot
public_inbound.dshield
public_inbound.dragonresearch
public_inbound.ciarmy
public_inbound.charleshaley
public_inbound.bruteforceblocker
public_inbound.botscout
public_inbound.blocklistde
public_inbound.autoshun
public_inbound.alienvault
Source (is contained)
% Overlap
0%
25%
50%
75%
100%

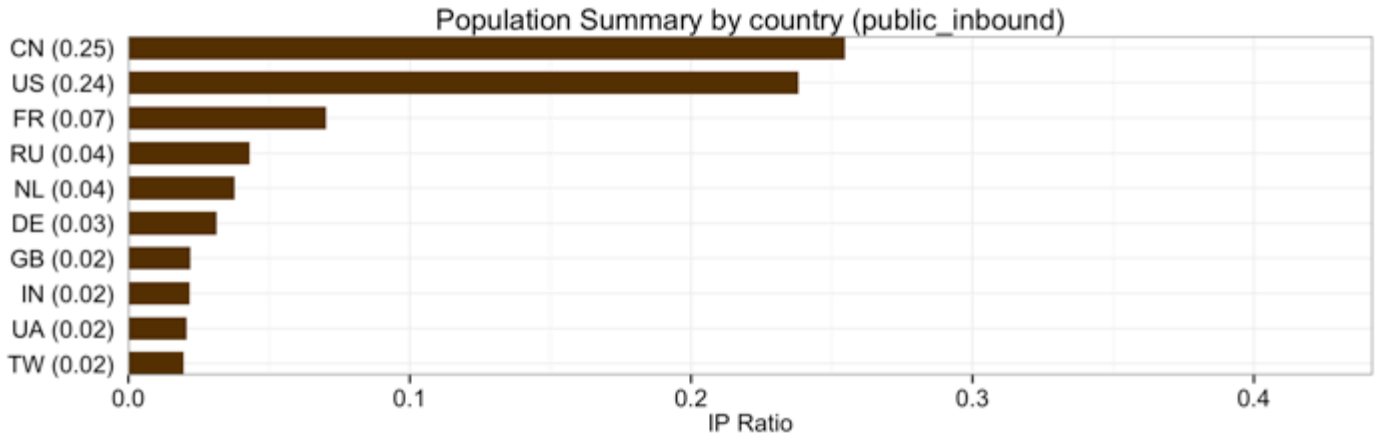

Overlap Test - Outbound Data - 20141101
Source (contains)
public_outbound.zeus
public_outbound.sslbl
public_outbound.spyeye
public_outbound.palevotracker
public_outbound.malwaregroup
public_outbound.malwaredomains
public_outbound.malwaredomainlist
public_outbound.malcode_zones
public_outbound.malcode
public_outbound.feodo
public_outbound.alienvault
Source (is contained)
% Overlap
0%
25%
50%
75%
100%

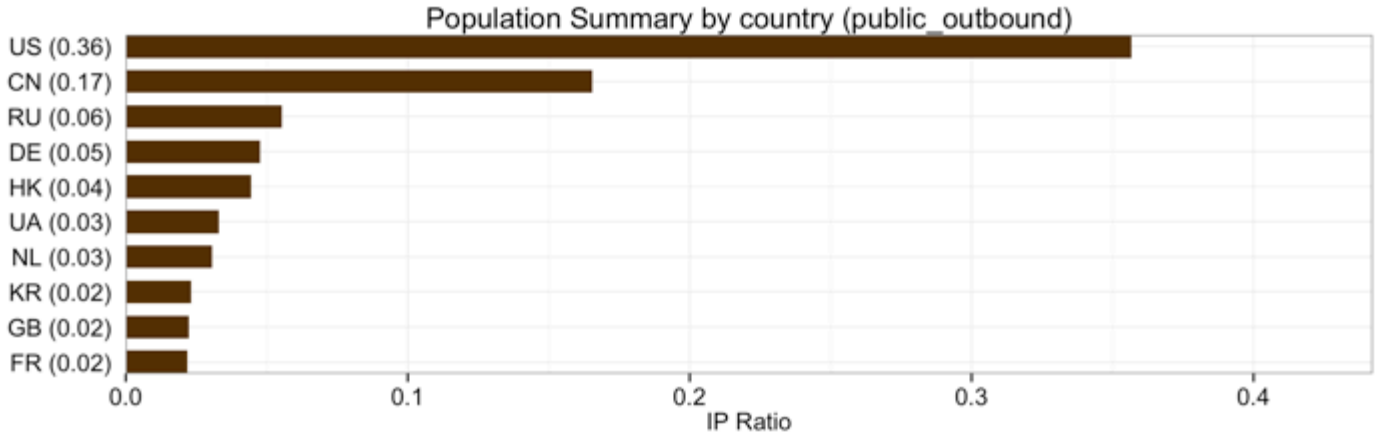
Population Test

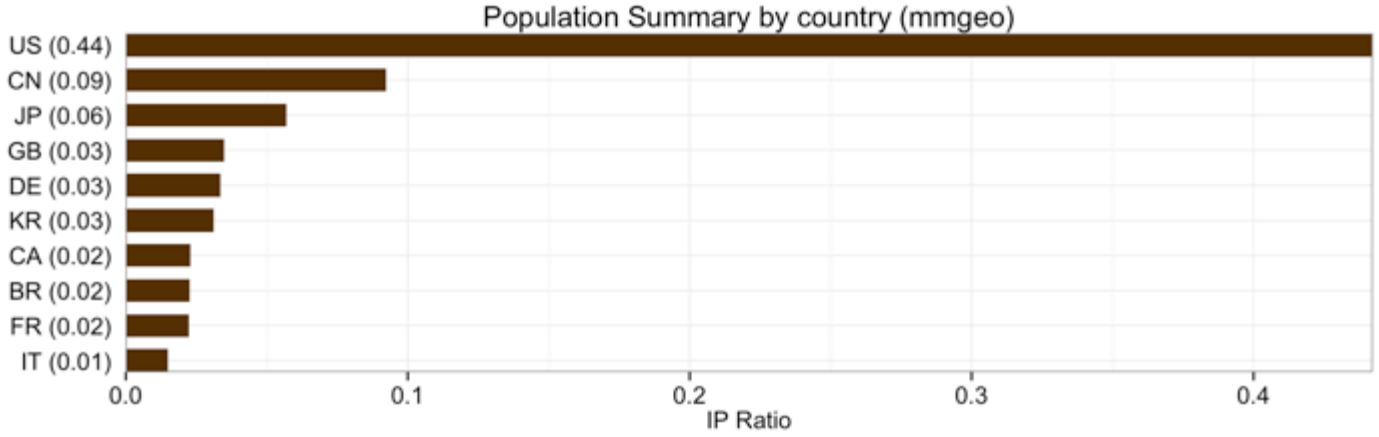
- Let us use the ASN and GeoIP databases that we used to enrich our data as a reference of the “true” population.

- But, but, human beings are unpredictable! We will never be able to forecast this!




Population Summary by country (public_inbound)
CN (0.25)
US (0.24)
FR (0.07)
RU (0.04)
NL (0.04)
DE (0.03)
GB (0.02)
IN (0.02)
UA (0.02)
TW (0.02)
0.0
0.1
0.2
0.3
0.4
IP Ratio


Population Summary by country (public_outbound)
US (0.36)
CN (0.17)
RU (0.06)
DE (0.05)
HK (0.04)
UA (0.03)
NL (0.03)
KR (0.02)
GB (0.02)
FR (0.02)
0.0
0.1
0.2
0.3
0.4
IP Ratio


Population Summary by country (mmgeo)
US (0.44)
CN (0.09)
JP (0.06)
GB (0.03)
DE (0.03)
KR (0.03)
CA (0.02)
BR (0.02)
FR (0.02)
IT (0.01)
0.0
0.1
0.2
0.3
0.4
IP Ratio

Is your sampling poll as random as you think?

Can we get a better look?

- Statistical inference-based comparison models (hypothesis testing)
  - Exact binomial tests (when we have the "true" pop)
  - Chi-squared proportion tests (similar to independence tests)

```
tests = tiq.test.populationInference(complete.pop$mmgeo,
                                     outbound.pop$public_outbound, "country",
                                     exact = TRUE, top=10)

# Whose proportion is bigger than it should be?
```

```
##    country conf.int.start conf.int.end    p.value
## 1:      TH       0.047044      0.05415  0.000e+00
## 2:      US       0.025335      0.04111  9.406e-17
## 5:      HK       0.014238      0.01868 2.412e-128
## 6:      NL       0.007818      0.01268  4.091e-23
```

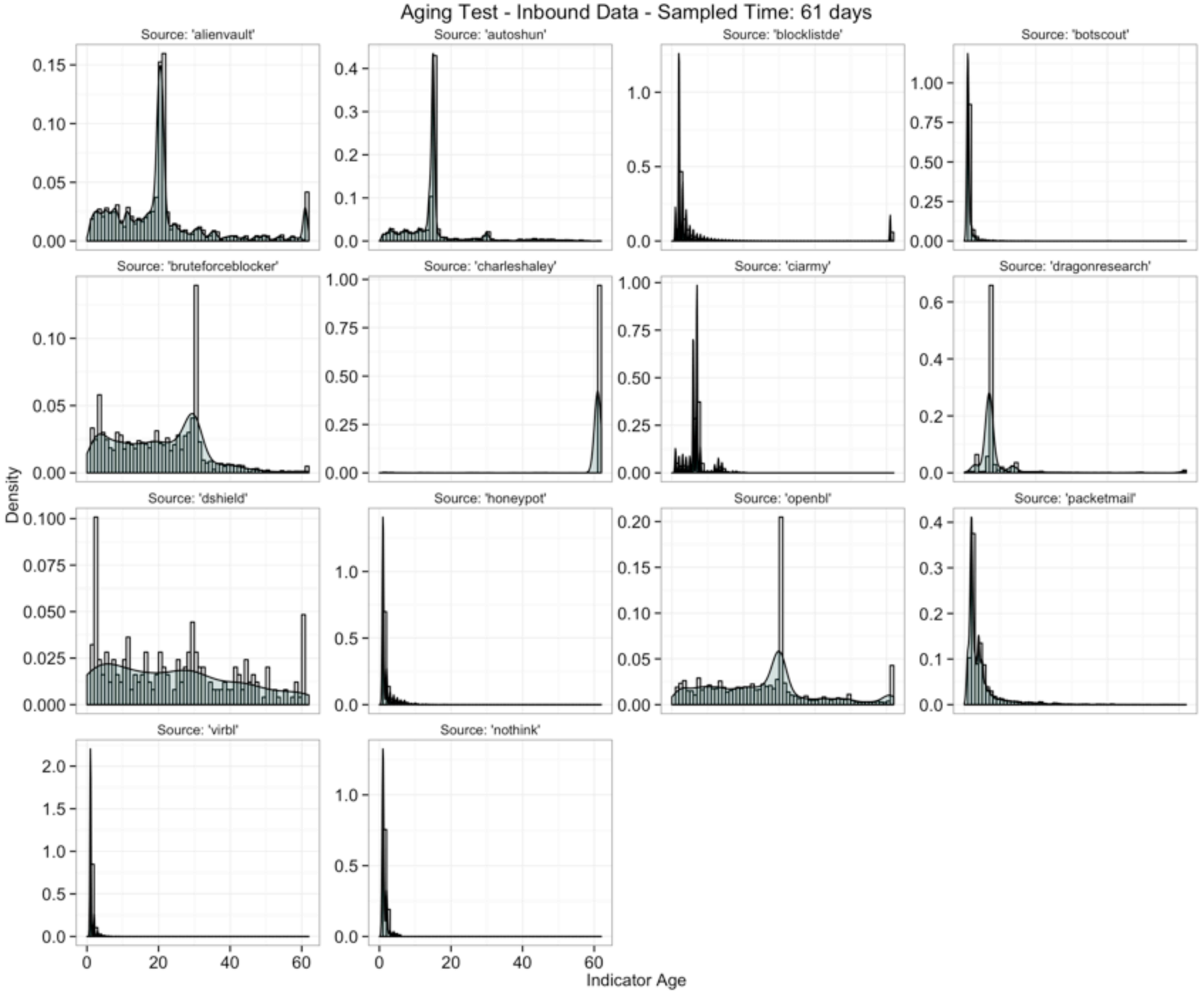
```
# Whose is smaller?
tests[p.value < 0.05/10 & conf.int.start < 0][order(conf.int.start, decreasing=F)]
```

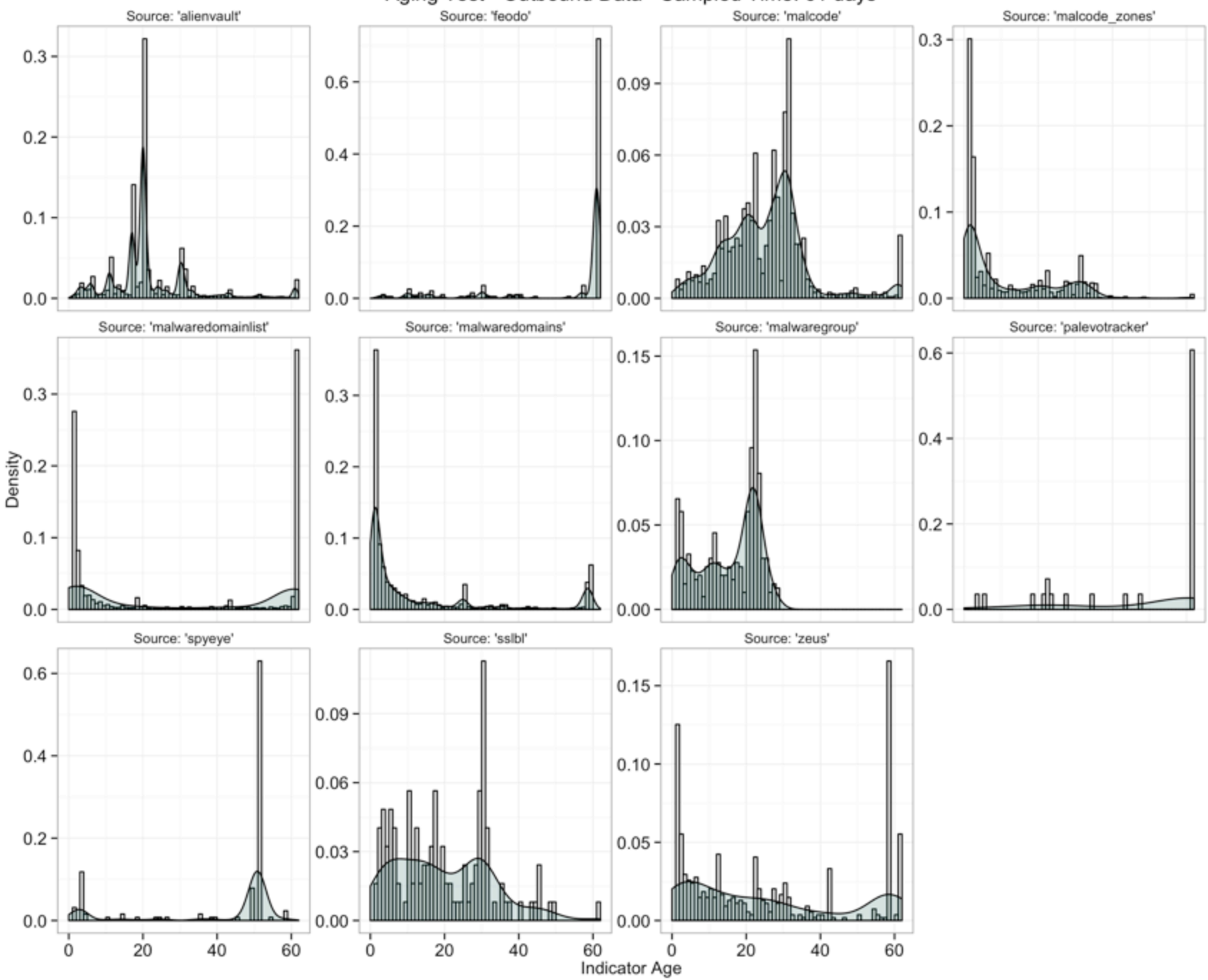
```
##    country conf.int.start conf.int.end   p.value
## 1:      GB      -0.01926    -0.015040 8.988e-38
## 2:      CN      -0.01469    -0.005996 5.356e-06
```

```
# And whose is the same? ¯\_(ツ)_/¯
tests[p.value > 0.05/10]
```

```
##    country conf.int.start conf.int.end p.value
## 1:      DE      -0.002366     0.003411  0.7553
```

Aging Test – Is someone cleaning this up eventually?

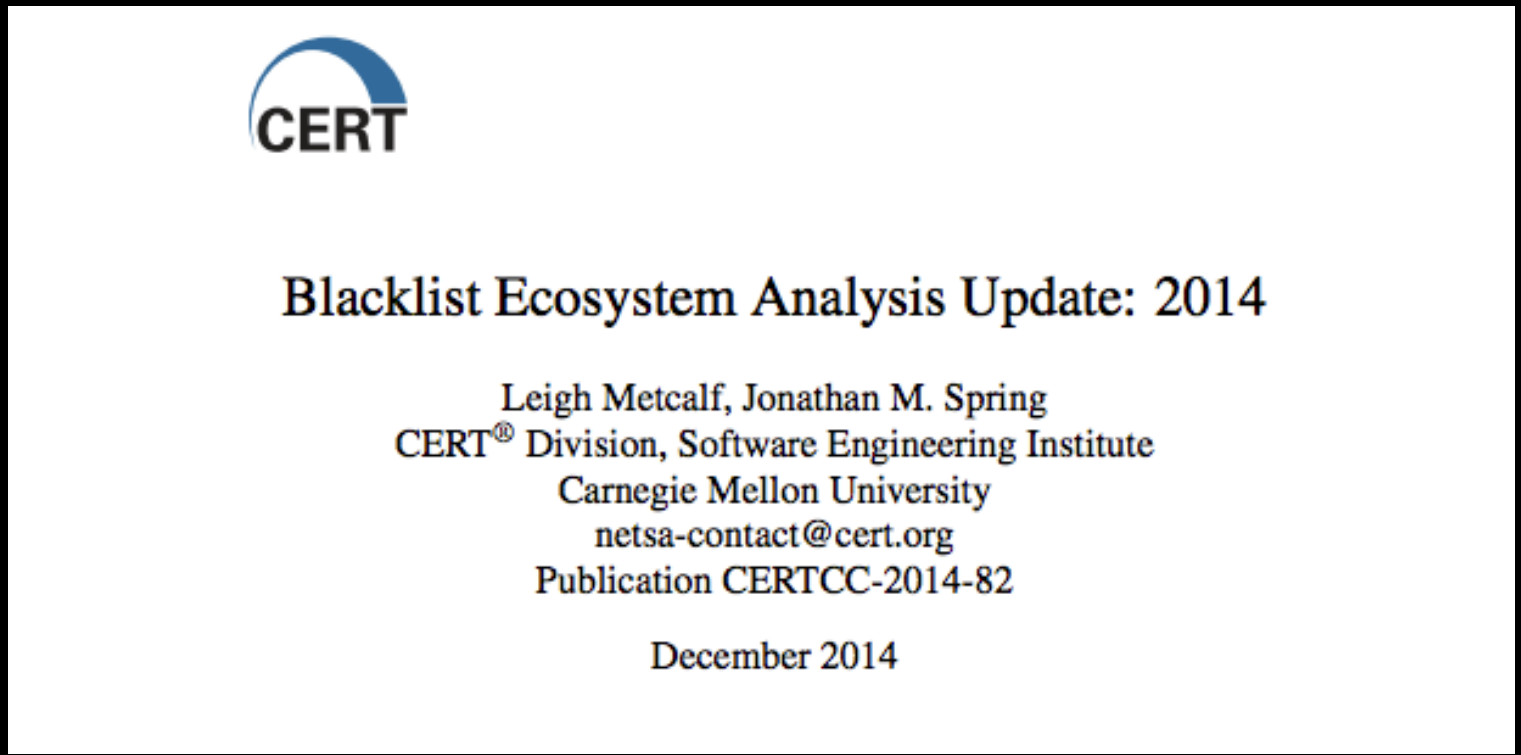

Aging Test - Inbound Data - Sampled Time: 61 days
Source: 'alienvault'
Source: 'autoshun'
Source: 'blocklistde'
Source: 'botscout'
Source: 'bruteforceblocker'
Source: 'charleshaley'
Source: 'ciarmy'
Source: 'dragonresearch'
Source: 'dshield'
Source: 'honeypot'
Source: 'openbl'
Source: 'packetmail'
Source: 'virbl'
Source: 'nothink'
Density
Indicator Age

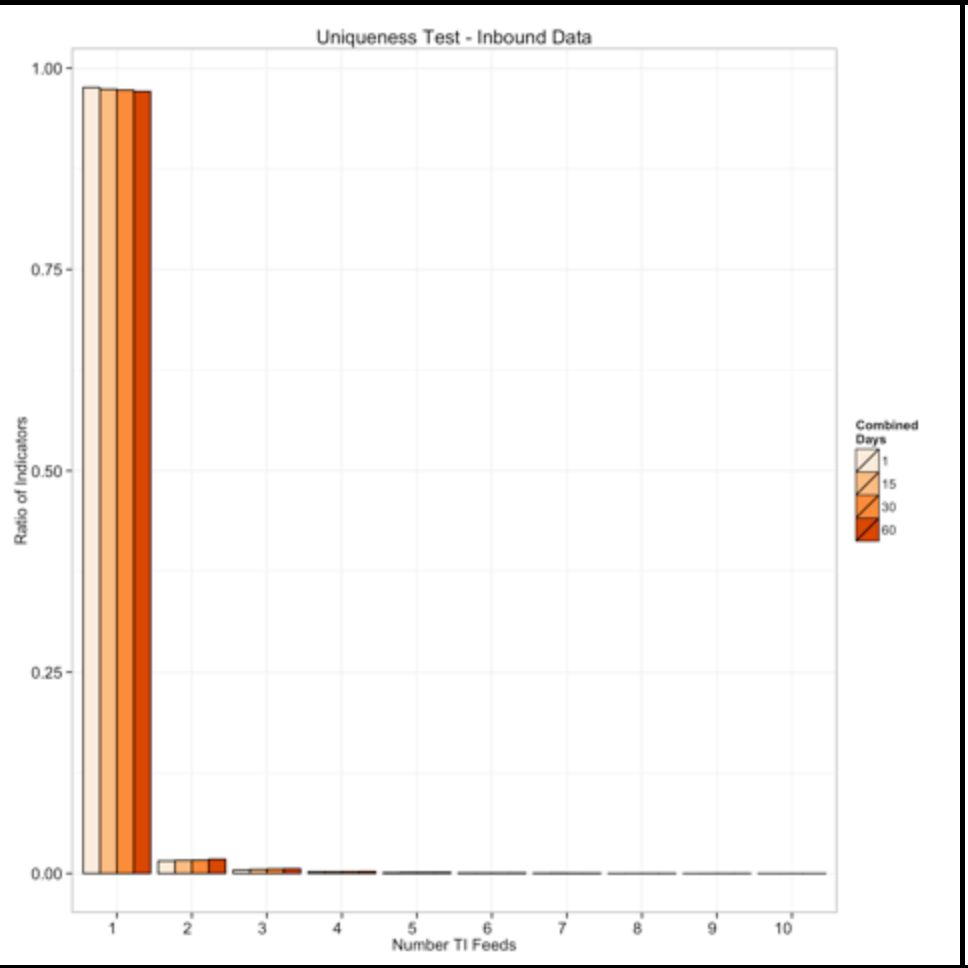

Aging Test - Outbound Data - Sampled Time: 61 days
Source: 'alienvault'
Source: 'feodo'
Source: 'malcode'
Source: 'malcode_zones'
Source: 'malwaredomainlist'
Source: 'malwaredomains'
Source: 'malwaregroup'
Source: 'palevotracker'
Source: 'spyeye'
Source: 'sslbl'
Source: 'zeus'
Density
Indicator Age
0
20
40
60

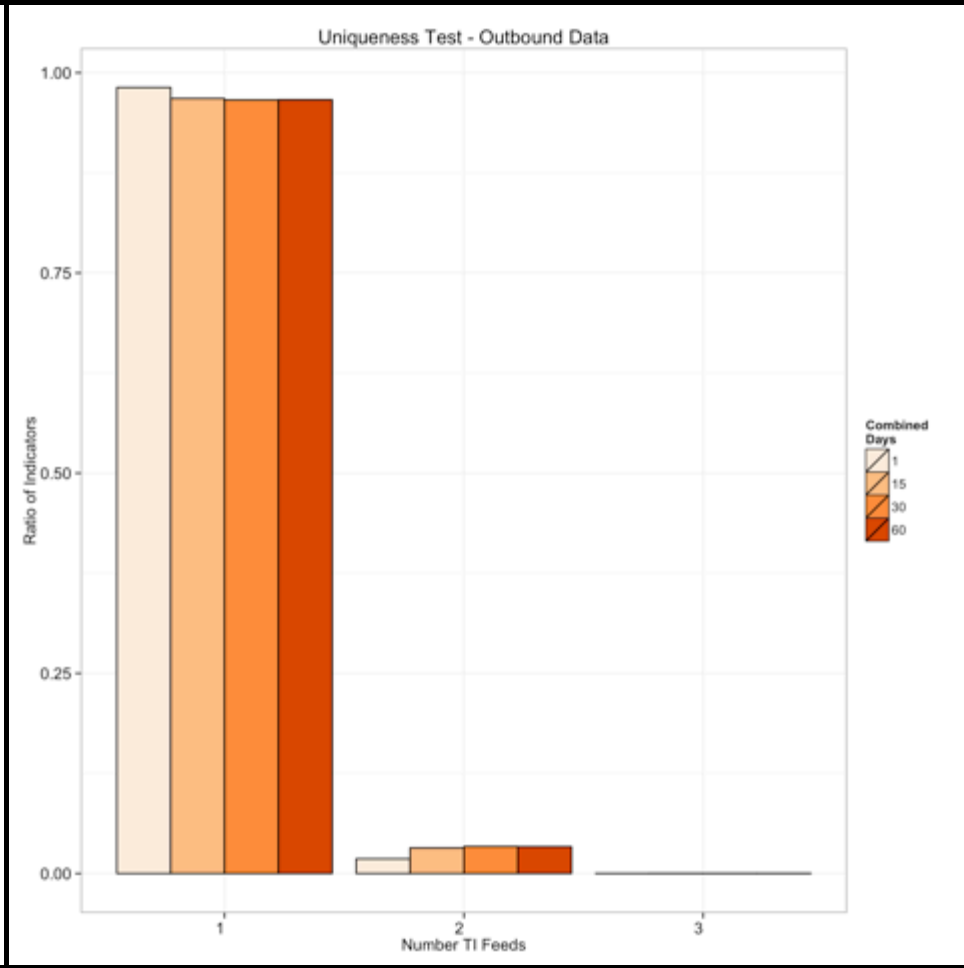
Uniqueness Test

Uniqueness Test

- “Domain-based indicators are unique to one list between 96.16% and 97.37%”
- “IP-based indicators are unique to one list between 82.46% and 95.24% of the time”








```
##    count     ratio days
## 1:     1 0.9759170    1
## 2:     1 0.9737684   15
## 3:     1 0.9727630   30
## 4:     1 0.9710713   60
```

```
##    count     ratio days
## 1:     1 0.9815816    1
## 2:     1 0.9678728   15
## 3:     1 0.9660537   30
## 4:     1 0.9663163   60
```

try some delicious and healthy snacks
at our concession stand

Intermission

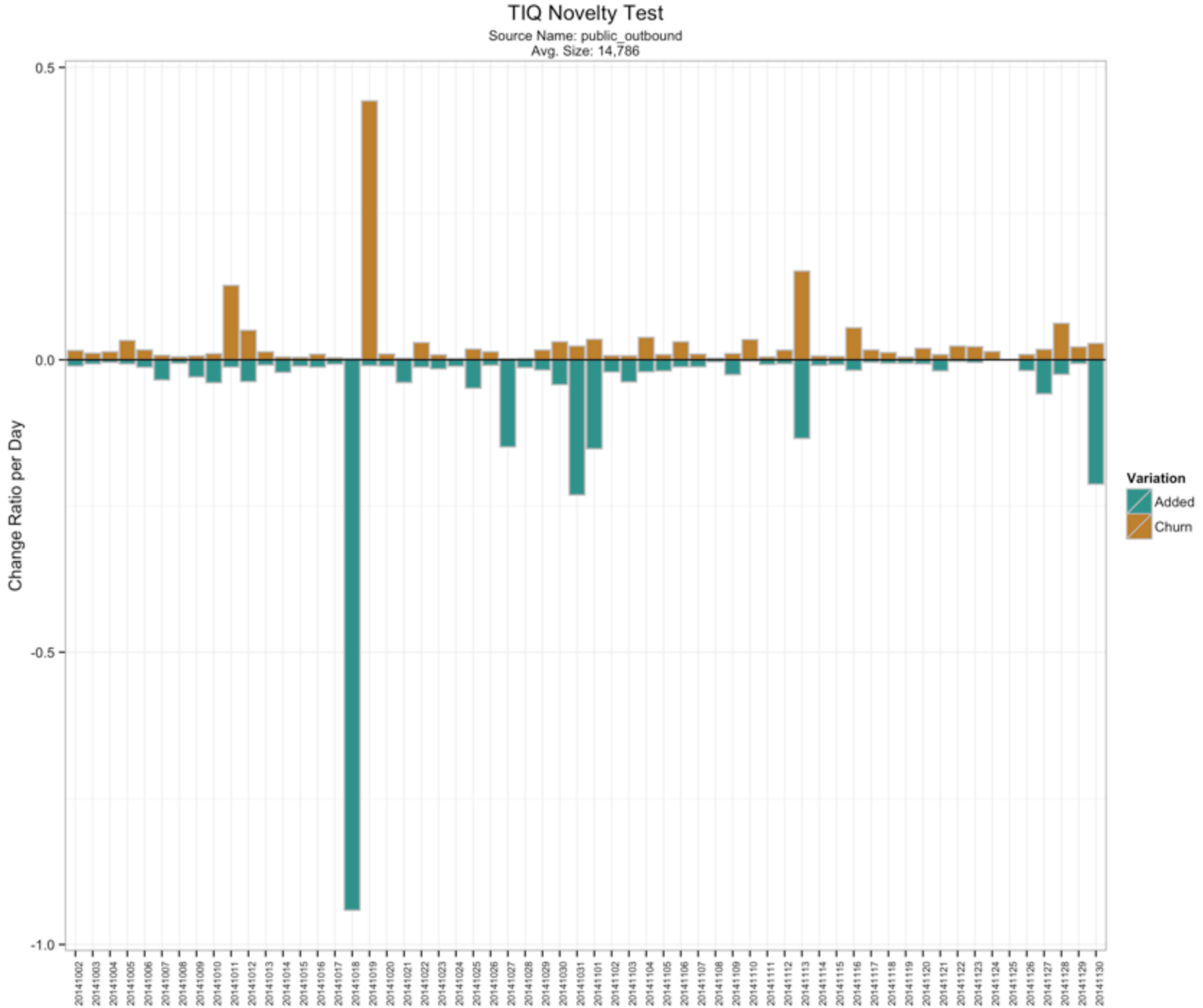
OPTION 1: Cool Story, Bro!

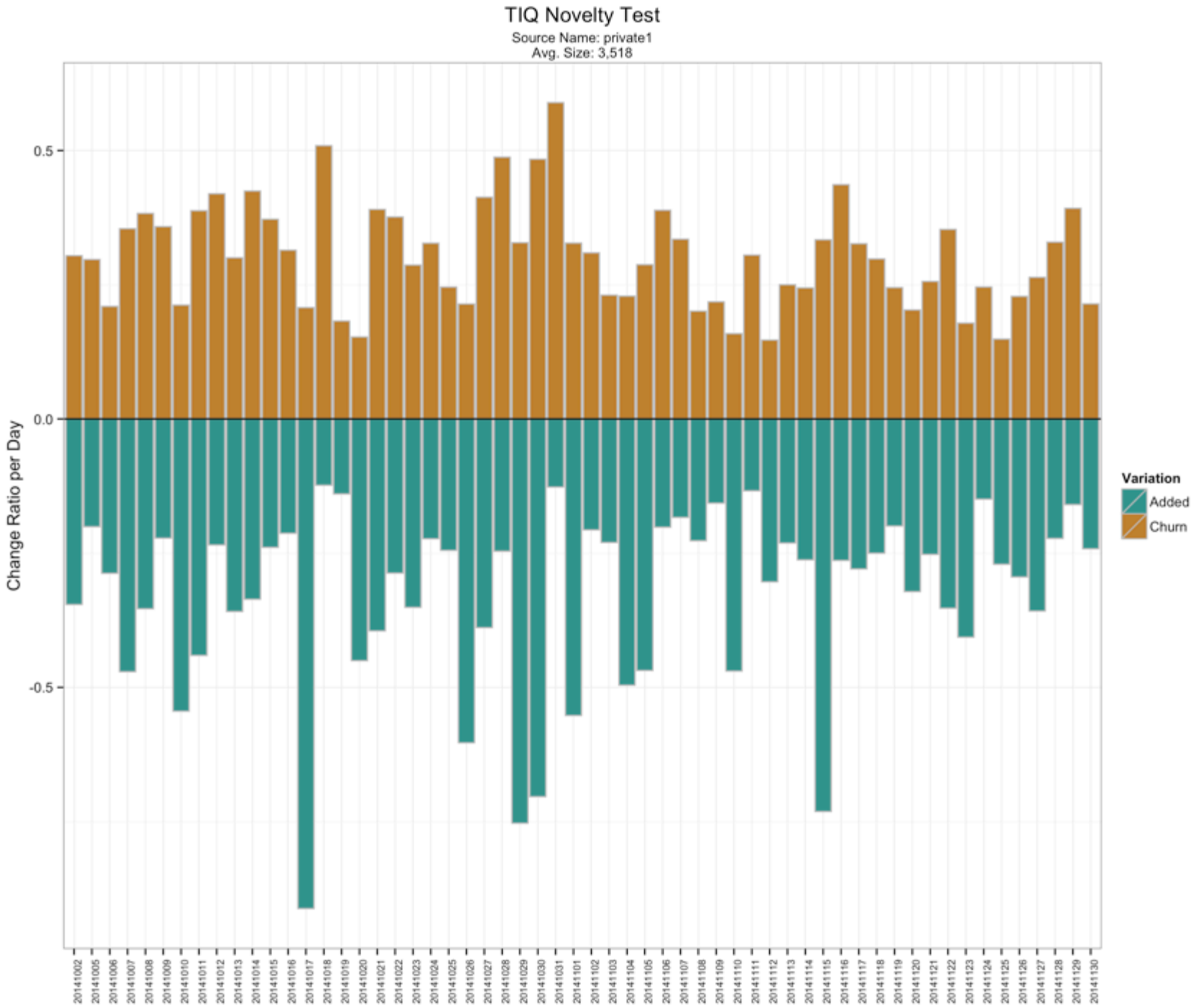
- Some of you are probably like:
  - "You Data Scientists and your algorithms, how quaint."
  - "Why aren't you doing some useful research like nation-state attribution?"

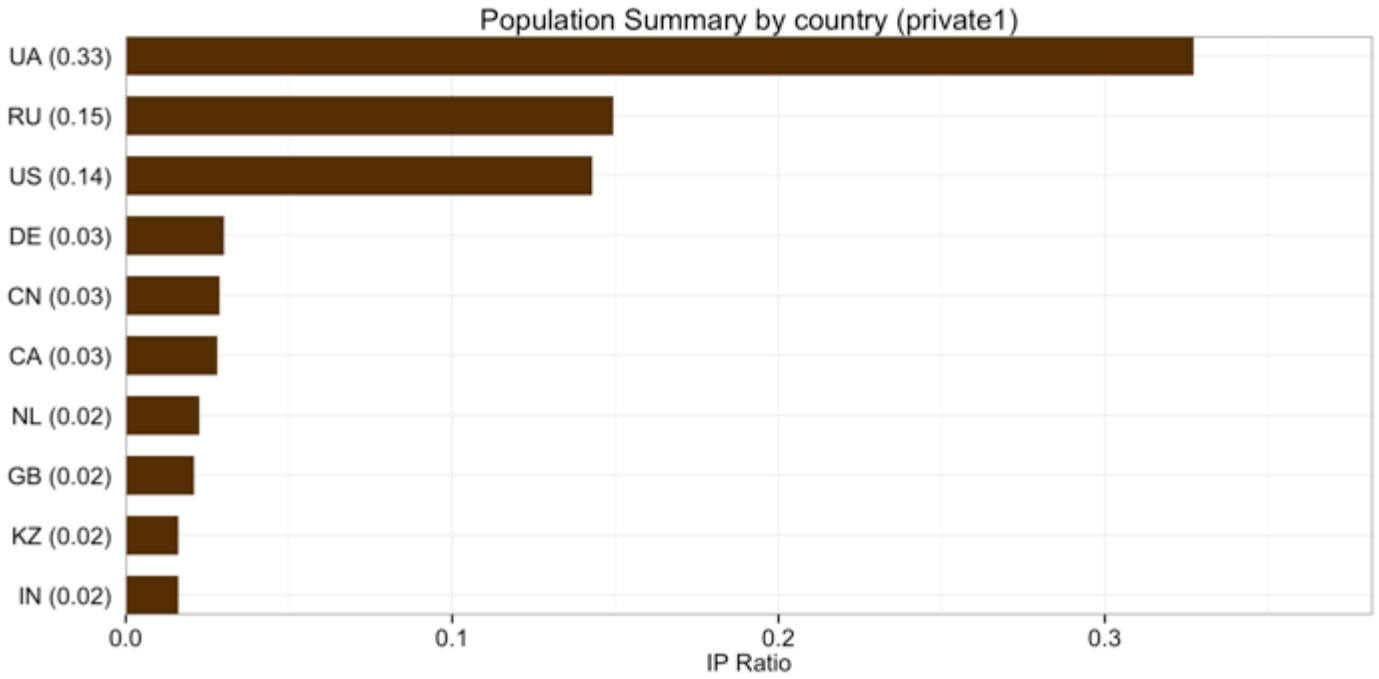
OPTION 2: How can I use this awesomeness on my data?

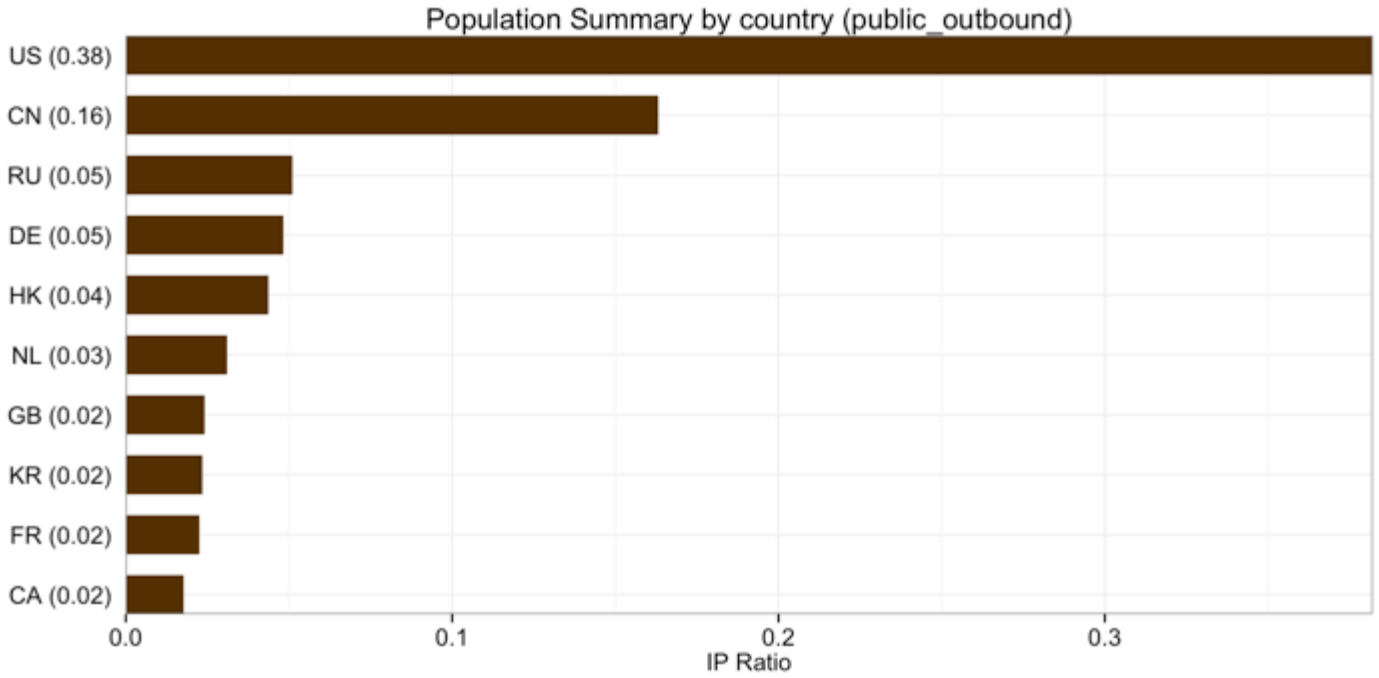
Use Case: Comparing Private Feeds

- How about using TIQ-TEST to evaluate a private intel feed?
- Trying stuff before you buy is usually a good idea. Just sayin'
- Let's compare a new feed, "private1", against our combined outbound indicators

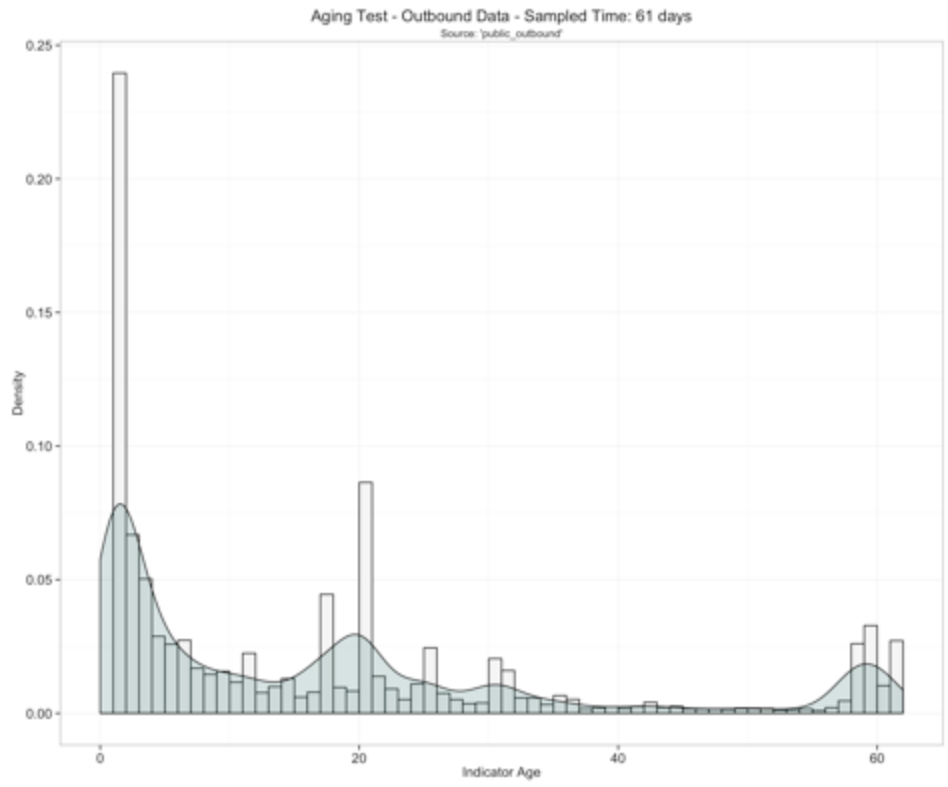

TIQ Novelty Test
Source Name: public_outbound
Avg. Size: 14,786
Change Ratio per Day
0.5
0.0
-0.5
-1.0
Variation
Added
Churn
20141002
20141003
20141004
20141005
20141006
20141007
20141008
20141009
20141010
20141011
20141012
20141013
20141014
20141015
20141016
20141017
20141018
20141019
20141020
20141021
20141022
20141023
20141024
20141025
20141026
20141027
20141028
20141029
20141030
20141031
20141101
20141102
20141103
20141104
20141105
20141106
20141107
20141108
20141109
20141110
20141111
20141112
20141113
20141114
20141115
20141116
20141117
20141118
20141119
20141120
20141121
20141122
20141123
20141124
20141125
20141126
20141127
20141128
20141129
20141130

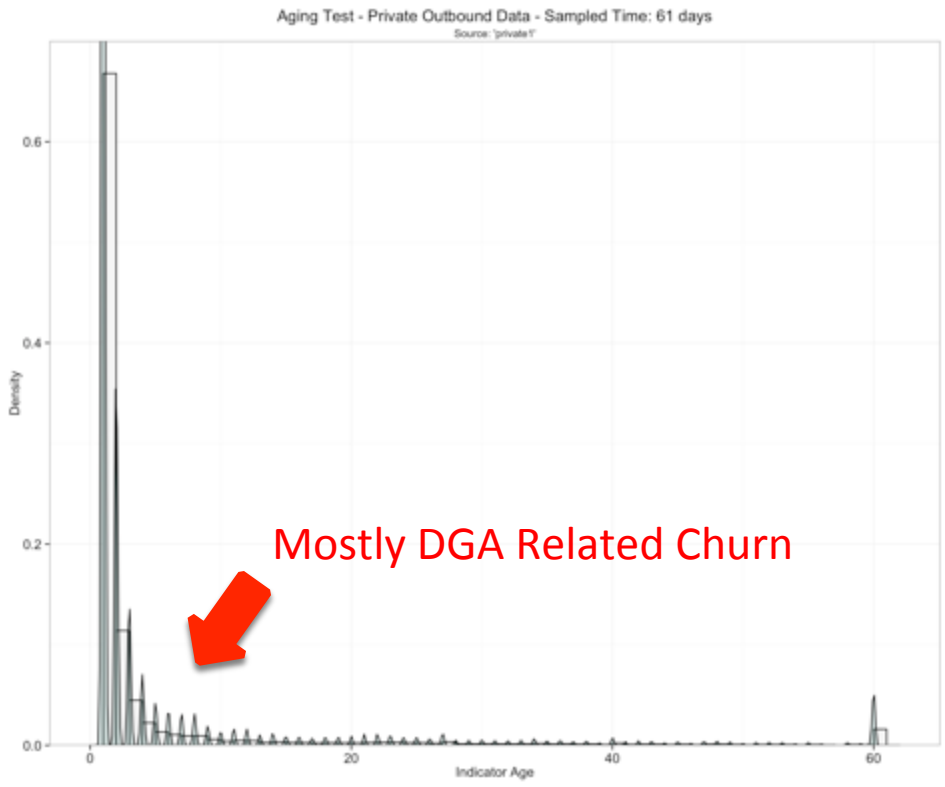

TIQ Novelty Test
Source Name: private1
Avg. Size: 3,518
Change Ratio per Day
0.5
0.0
-0.5
Variation
Added
Churn
20141002
20141005
20141006
20141007
20141008
20141009
20141010
20141011
20141012
20141013
20141014
20141015
20141016
20141017
20141018
20141019
20141020
20141021
20141022
20141023
20141024
20141025
20141026
20141027
20141028
20141029
20141030
20141031
20141101
20141102
20141103
20141104
20141105
20141106
20141107
20141108
20141109
20141110
20141111
20141112
20141113
20141114
20141115
20141116
20141117
20141118
20141119
20141120
20141121
20141122
20141123
20141124
20141125
20141126
20141127
20141128
20141129
20141130


Population Summary by country (private1)
UA (0.33)
RU (0.15)
US (0.14)
DE (0.03)
CN (0.03)
CA (0.03)
NL (0.02)
GB (0.02)
KZ (0.02)
IN (0.02)
0.0
0.1
0.2
0.3
IP Ratio

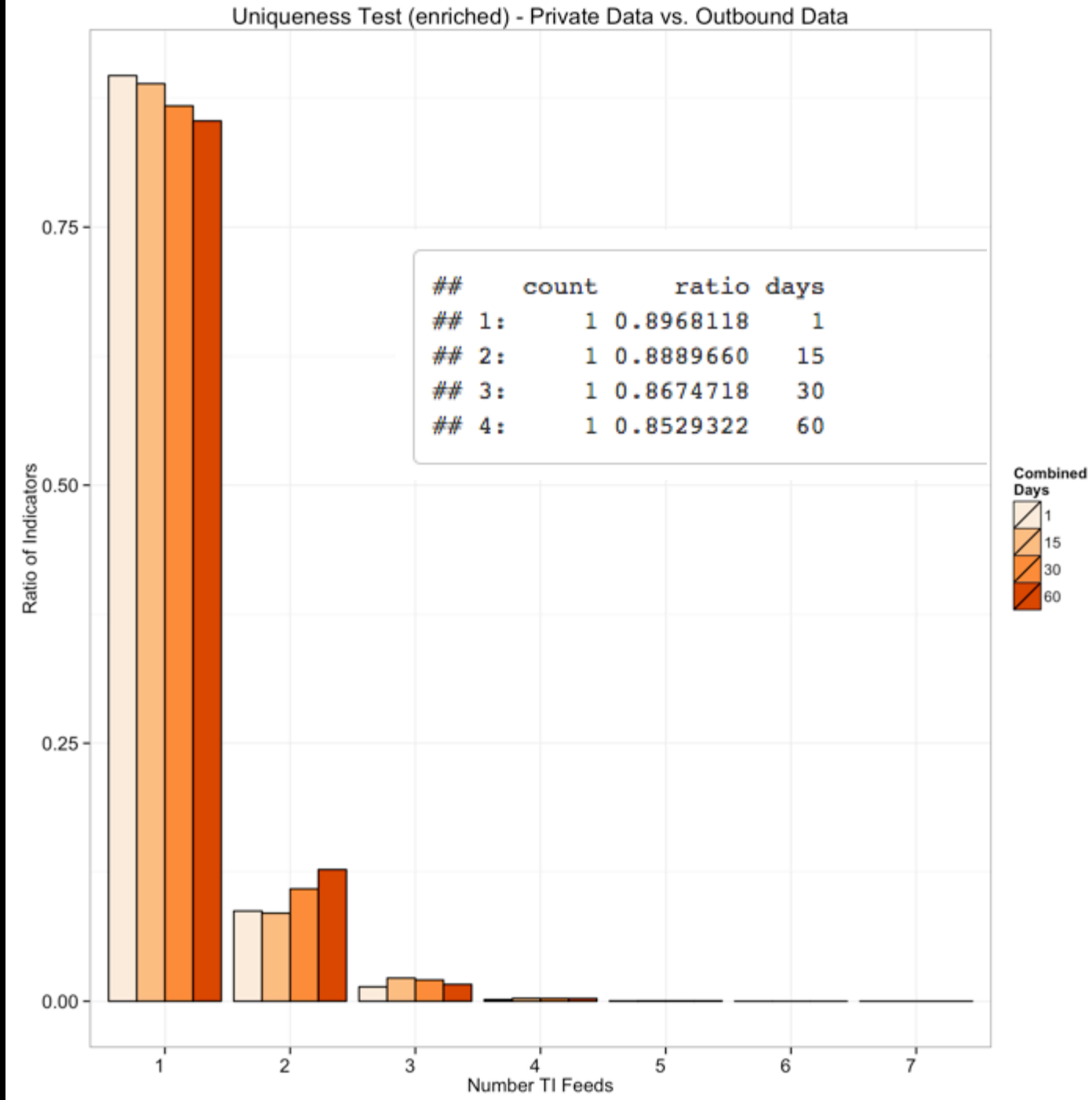

Population Summary by country (public_outbound)
US (0.38)
CN (0.16)
RU (0.05)
DE (0.05)
HK (0.04)
NL (0.03)
GB (0.02)
KR (0.02)
FR (0.02)
CA (0.02)
0.0
0.1
0.2
0.3
IP Ratio

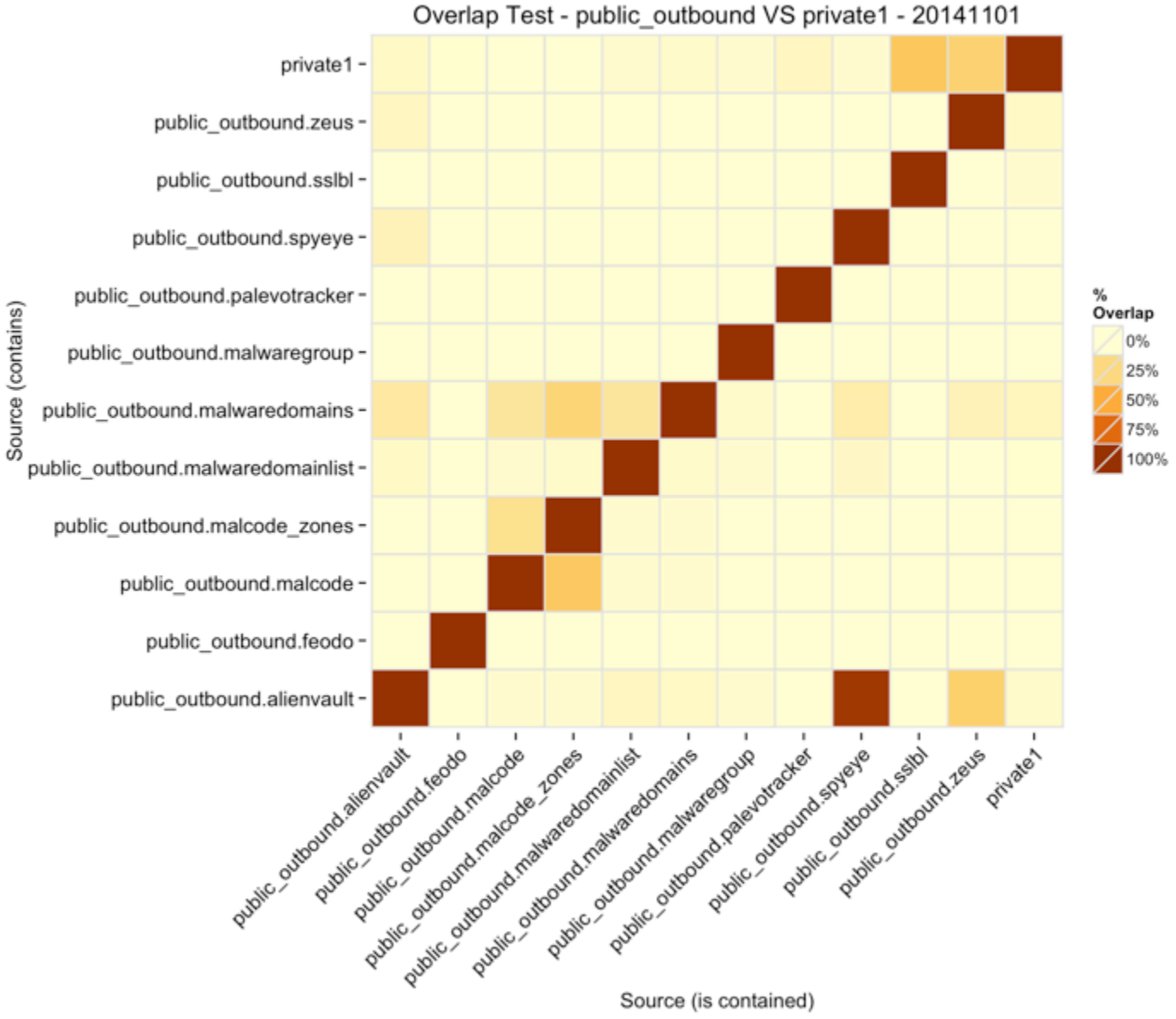
Aging Test





- I guess most DGAs rotate every 24 hours, right?
- Rotation means the private data is still “fresh”, from research or DGA generation procedures


Uniqueness Test (enriched) - Private Data vs. Outbound Data
count ratio days
1: 1 0.8968118 1
2: 1 0.8889660 15
3: 1 0.8674718 30
4: 1 0.8529322 60
Ratio of Indicators
0.00
0.25
0.50
0.75
1
2
3
4
5
6
7
Number TI Feeds
Combined Days
1
15
30
60


Overlap Test - public_outbound VS private1 - 20141101
Source (contains)
private1
public_outbound.zeus
public_outbound.sslbl
public_outbound.spyeye
public_outbound.palevotracker
public_outbound.malwaregroup
public_outbound.malwaredomains
public_outbound.malwaredomainlist
public_outbound.malcode_zones
public_outbound.malcode
public_outbound.feodo
public_outbound.alienvault
Source (is contained)
public_outbound.alienvault
public_outbound.feodo
public_outbound.malcode
public_outbound.malcode_zones
public_outbound.malwaredomainlist
public_outbound.malwaredomains
public_outbound.malwaregroup
public_outbound.palevotracker
public_outbound.spyeye
public_outbound.sslbl
public_outbound.zeus
private1
% Overlap
0%
25%
50%
75%
100%

1.

A++

1. Relatively poor eBay feedback, compared to the community-accepted standard of **A+++++++**. Often used for negative feedback responses or as revenge.

2. A++ is also a programming language rumored to be used by more than three people worldwide.

"After getting my address from eBay, this seller broke into my house and killed my pets. A++"

"Man, this buyer sucks! I'll get him back by leaving him a mere A++ feedback."

by Jin64 September 02, 2007

A++ WOULD THREAT INTEL AGAIN

Take Away

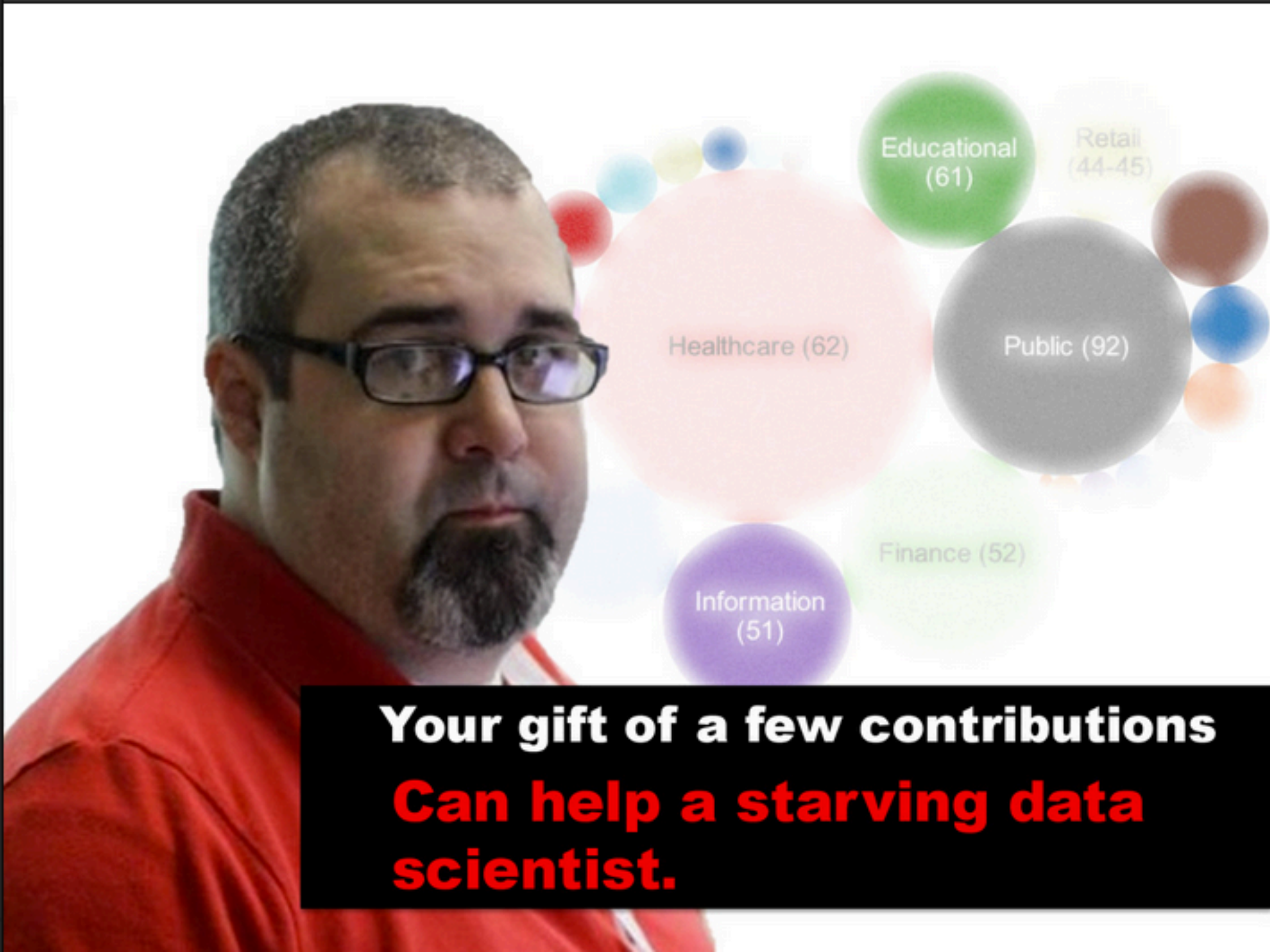
- Analyze your data. Extract value from it!
- Try before you buy! Different test results mean different things to different orgs.

- Try the sample data, replicate the experiments:
  - https://github.com/mlsecproject/tiq-test-Winter2015
  - http://rpubs.com/alexcpsec/tiq-test-Winter2015

Greets and Thanks!

- @kafeine and John Bambenek for access to their feeds
- @kylemaxwell, @paul4pc and all contributors for COMBINE
- @bfist for his work on http://sony.attributed.to
- @hrbrmstr for IPEW and contribs for TIQ-TEST

- All the MLSec Project community peeps!


Educational (61)
Retail (44-45)
Healthcare (62)
Public (92)
Finance (52)
Information (51)
Your gift of a few contributions
Can help a starving data scientist.

MLSec
MACHINE LEARNING SECURITY

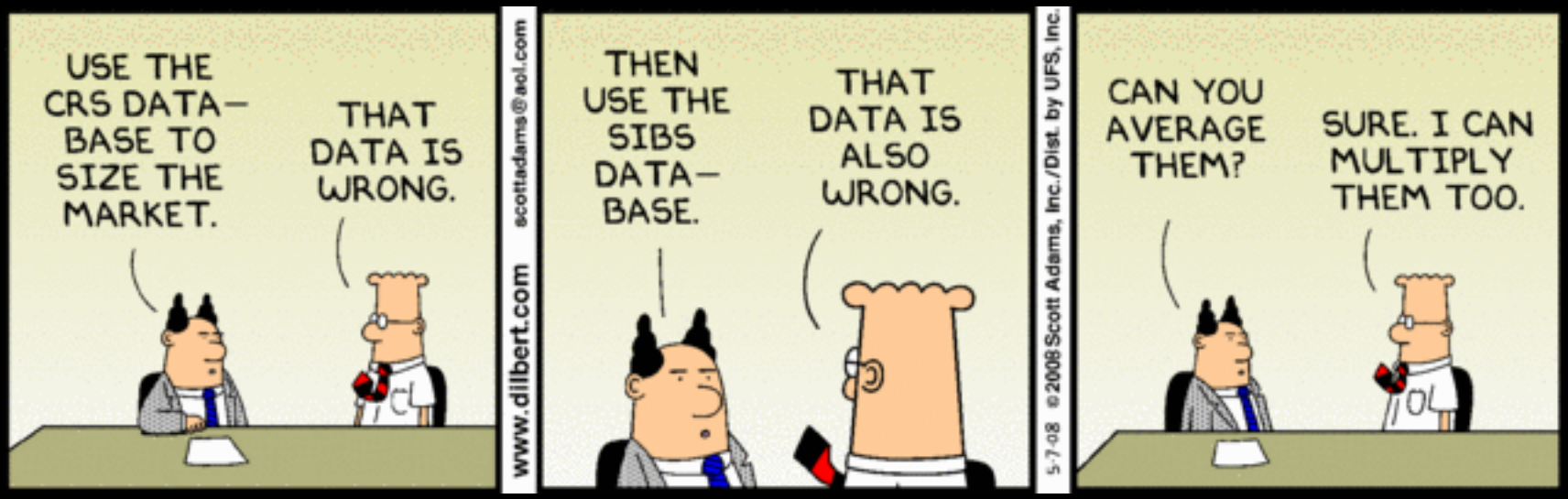
Thanks!

NIDDEL™

- Q&A?
- Feedback!

Alex Pinto
@alexcpsec
@MLSecProject
@NiddelCorp

Alexandre Sieira
@AlexandreSieira
@MLSecProject
@NiddelCorp



”The measure of intelligence is the ability to change."

- Albert Einstein