Measuring the IQ of your Threat Intelligence Feeds (#TIQtest)

Alex Pinto
MLSec Project
@alexcpsec
@MLSecProject

Kyle Maxwell
Researcher
@kylemaxwell

whoami(s)

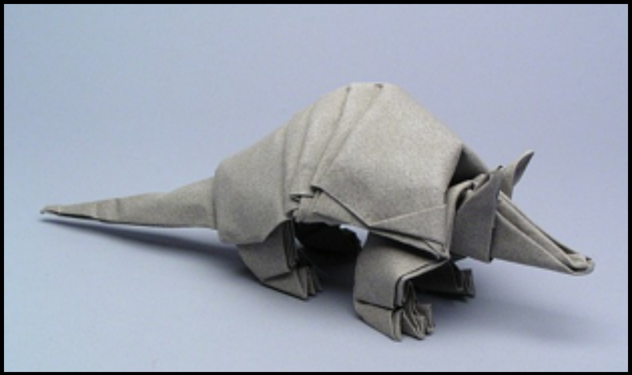
Alex Pinto

- Science guy at MLSec Project
- ML trainer
- Network security aficionado
- Tortured by SIEMs as a child
- Hacker Spirit Animal™: CAFFEINATED CAPYBARA

(https://secure.flickr.com/photos/kobashi_san/)

Kyle Maxwell

- Researcher at [REDACTED]
- Math Smuggler
- Recovering Incident Responder
- GPL zealot
- Hacker Spirit Animal™: AXIOMATIC ARMADILLO



(http://www.langorigami.com/art/gallery/gallery.php?tag=mammals&name=armadillo)

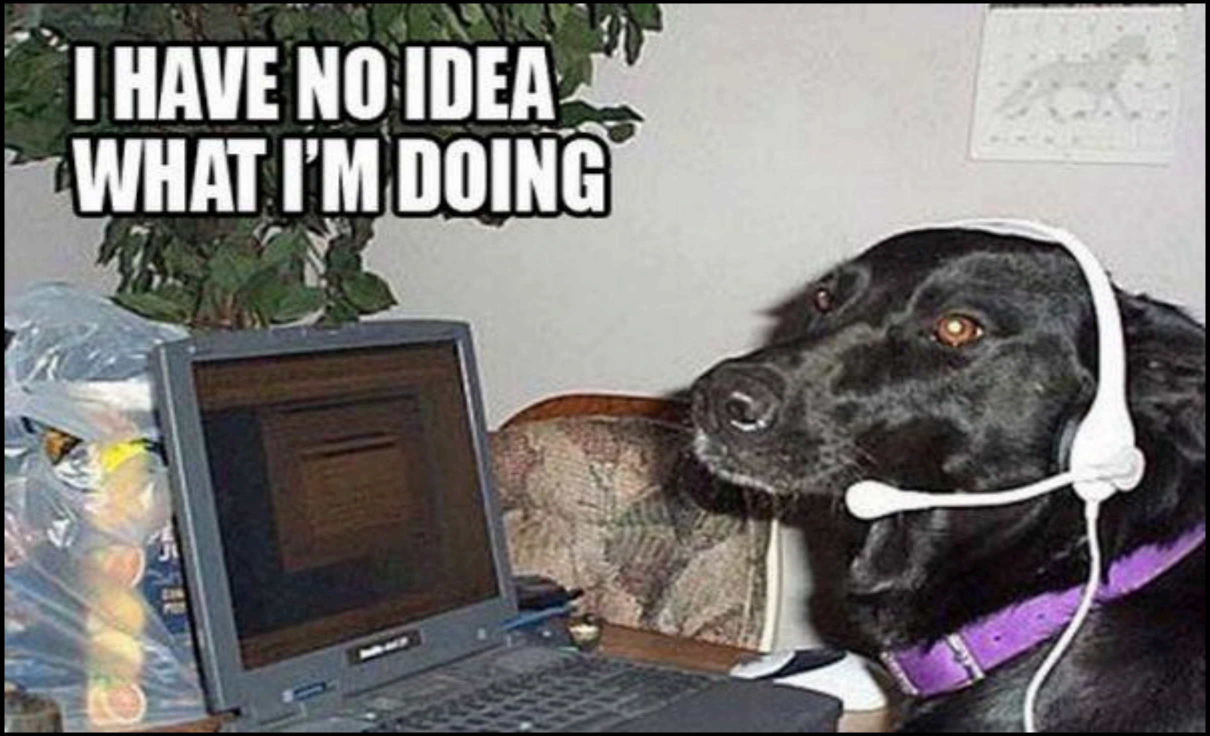
Agenda

- Threat Intel 102
- Measuring Intelligence
- Data Preparation
- Testing the Data
- Tools:
  - COMBINE
  - TIQ-TEST
- Some parting ideas

(http://www.savagechickens.com/2008/12/iq-test.html)

Threat Intel 102: Capability and Intent

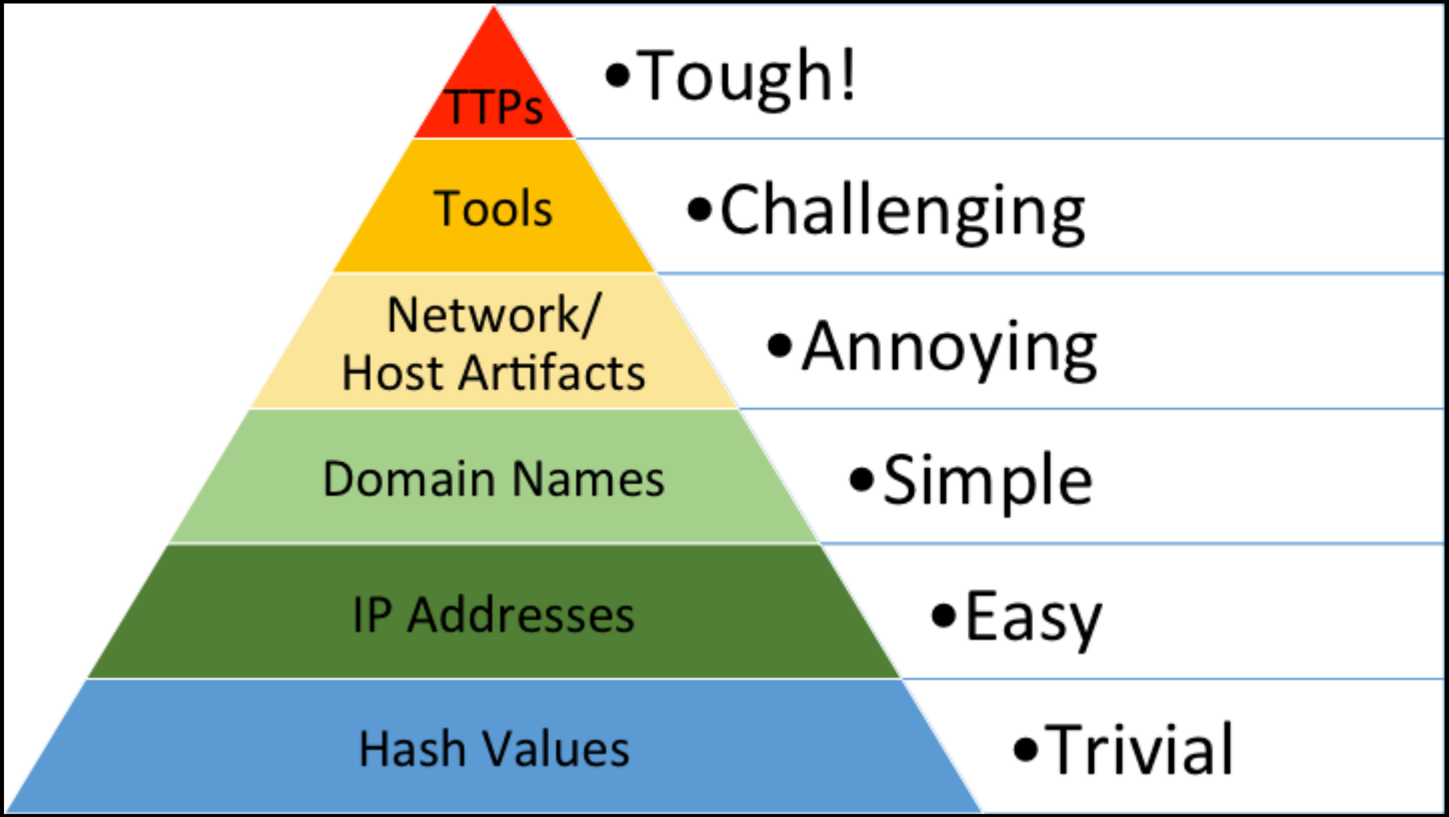
- What are they **able** to do?
- What are they **intending** to do?



Threat Intel 102: Cage Matches

- Signatures vs Indicators
- Data vs Intelligence
- Tactical vs Strategic
- Atomic vs Composite

Threat Intel 102: Pyramid of Pain

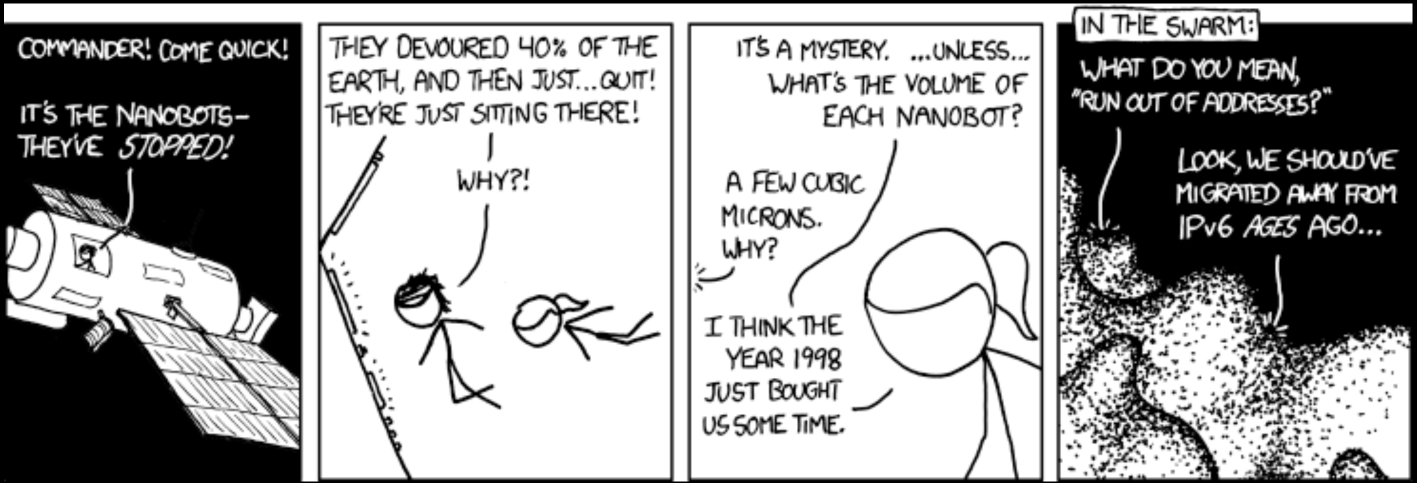


(David Bianco – Pyramid of Pain)

“Simple” and “easy” aren’t always

What about IP addresses?

- Approximately same value as hostnames (APT vs DGA)
- Finite resource (until IPv6, that is)
  - Managed / controlled by orgs
  - Difficulty / economic incentives / implied “cost”
  - Also, recyclable


COMMANDER! COME QUICK!
IT'S THE NANOBOTS– THEY'VE STOPPED!
THEY DEVOURED 40% OF THE EARTH, AND THEN JUST...QUIT! THEY'RE JUST SITTING THERE!
WHY?!
IT'S A MYSTERY. ...UNLESS... WHAT'S THE VOLUME OF EACH NANOBOT?
A FEW CUBIC MICRONS. WHY?
I THINK THE YEAR 1998 JUST BOUGHT US SOME TIME.
IN THE SWARM:
WHAT DO YOU MEAN, "RUN OUT OF ADDRESSES?"
LOOK, WE SHOULD'VE MIGRATED AWAY FROM IPv6 AGES AGO...

(https://xkcd.com/865/)

Given IP addresses harvested from TI feeds, can we measure how much they "help" our defense metrics?

Introducing TIQ-TEST

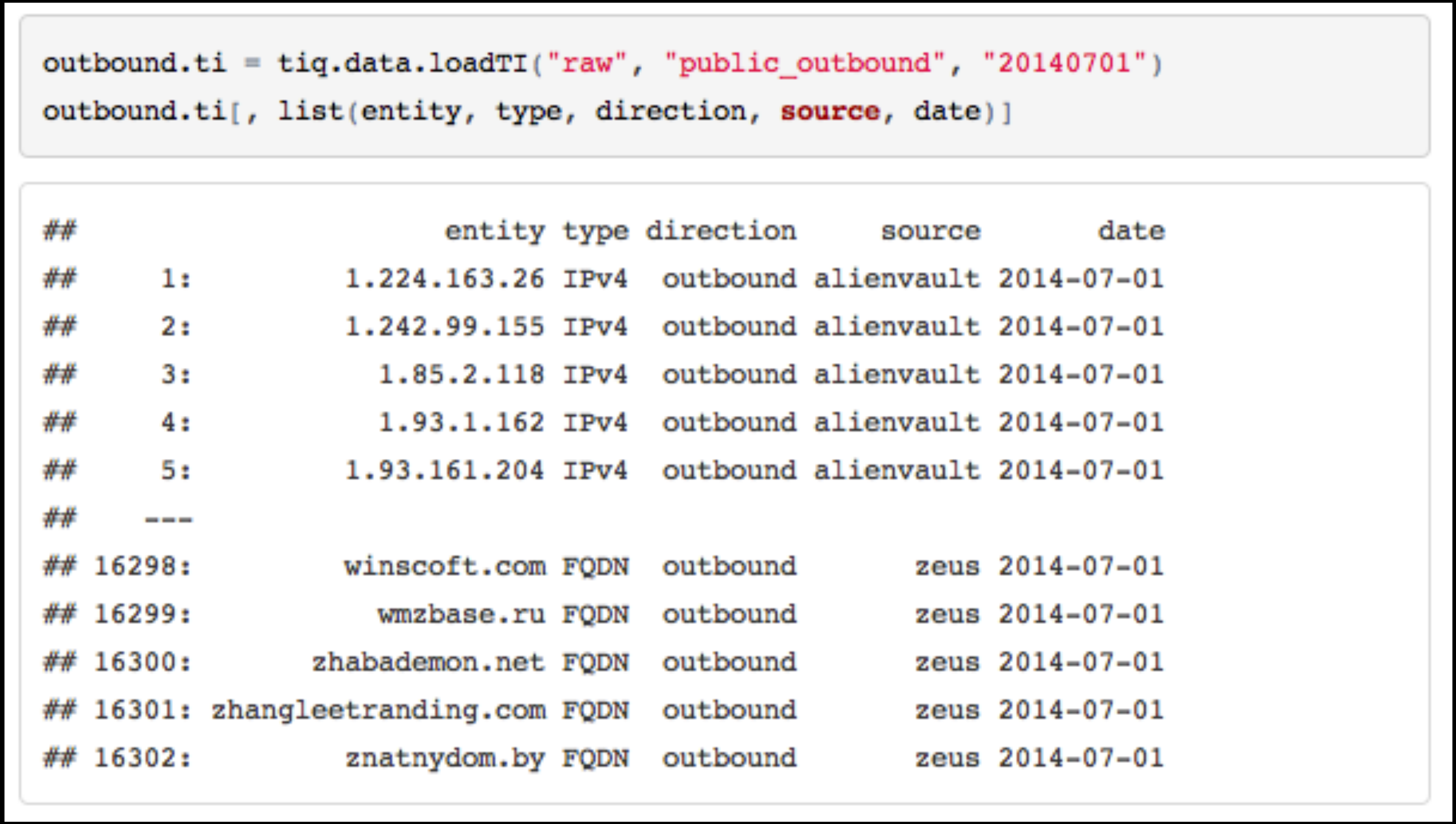
- All these tests are available as R functions at
  - https://github.com/mlsecproject/tiq-test
  - Have fun, prove me wrong, suggest stuff

- Tools that implement those tests
- Sample data + R Markdown file

- The excuse to learn a statistical language you were waiting for!

Data Sources – Types of data

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

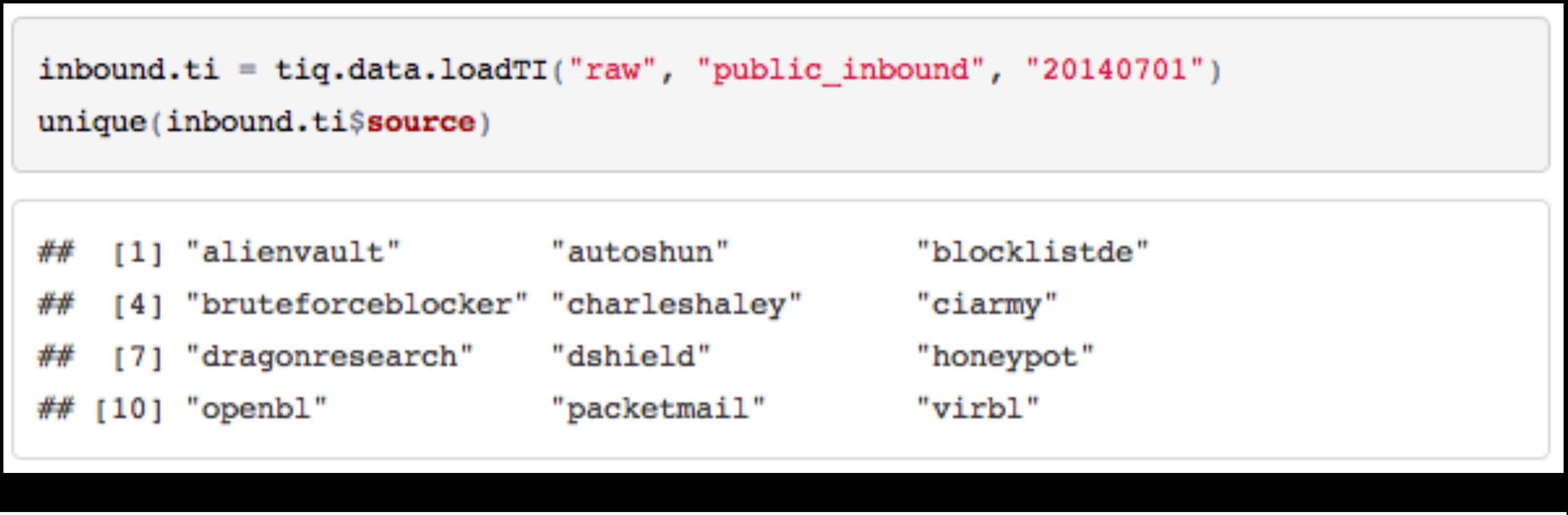


```
outbound.ti = tiq.data.loadTI("raw", "public_outbound", "20140701")
outbound.ti[, list(entity, type, direction, source, date)]
```

```
##                       entity type direction     source       date
##     1:          1.224.163.26 IPv4  outbound alienvault 2014-07-01
##     2:          1.242.99.155 IPv4  outbound alienvault 2014-07-01
##     3:            1.85.2.118 IPv4  outbound alienvault 2014-07-01
##     4:            1.93.1.162 IPv4  outbound alienvault 2014-07-01
##     5:          1.93.161.204 IPv4  outbound alienvault 2014-07-01
##    ---
## 16298:          winscoft.com FQDN  outbound       zeus 2014-07-01
## 16299:            wmzbase.ru FQDN  outbound       zeus 2014-07-01
## 16300:        zhabademon.net FQDN  outbound       zeus 2014-07-01
## 16301:  zhangleetranding.com FQDN  outbound       zeus 2014-07-01
## 16302:          znatnydom.by FQDN  outbound       zeus 2014-07-01
```

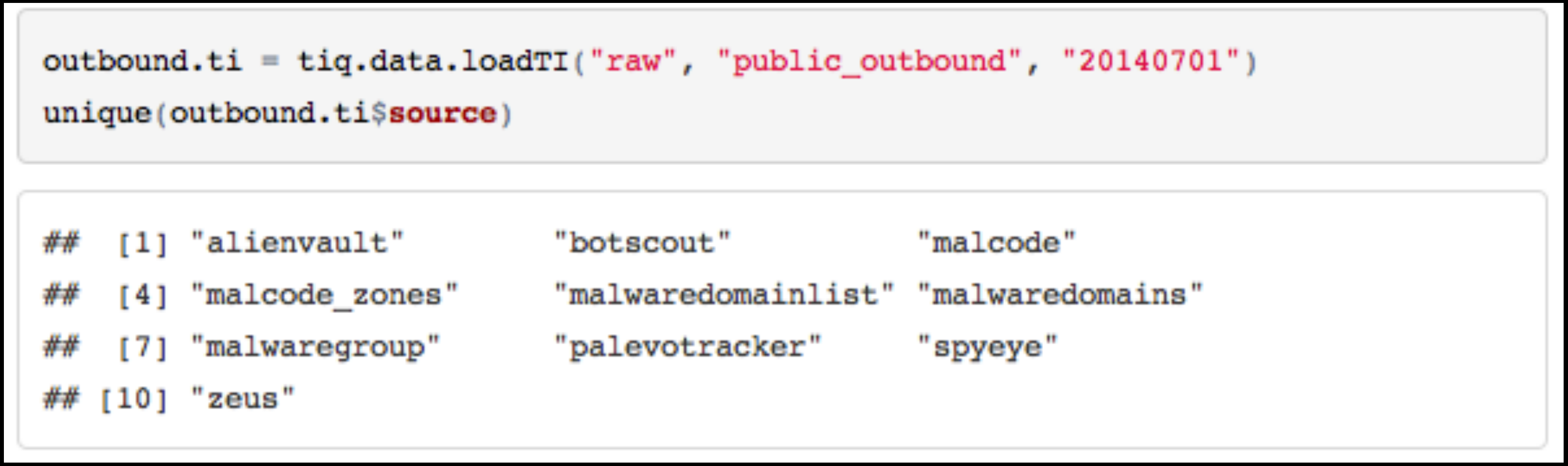
Data Sources – Feeds Selected

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



```
inbound.ti = tiq.data.loadTI("raw", "public_inbound", "20140701")
unique(inbound.ti$source)
```

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



```
outbound.ti = tiq.data.loadTI("raw", "public_outbound", "20140701")
unique(outbound.ti$source)
```

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

Data Preparation and Cleansing

- Convert the hostname data to IP addresses:
    - Active IP addresses for the respective date ("A" query)
    - Passive DNS from Farsight Security (DNSDB) 👍
- We removed non-public IPs from the dataset (RFC1918)
    - Yeah, we know it is a "parking technique"

BUT WHEN SHE TRACED THE KILLER'S IP ADDRESS... IT WAS IN THE 192.168/16 BLOCK!
GASP!

(https://xkcd.com/742/)

Data Preparation and Cleansing

- 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"

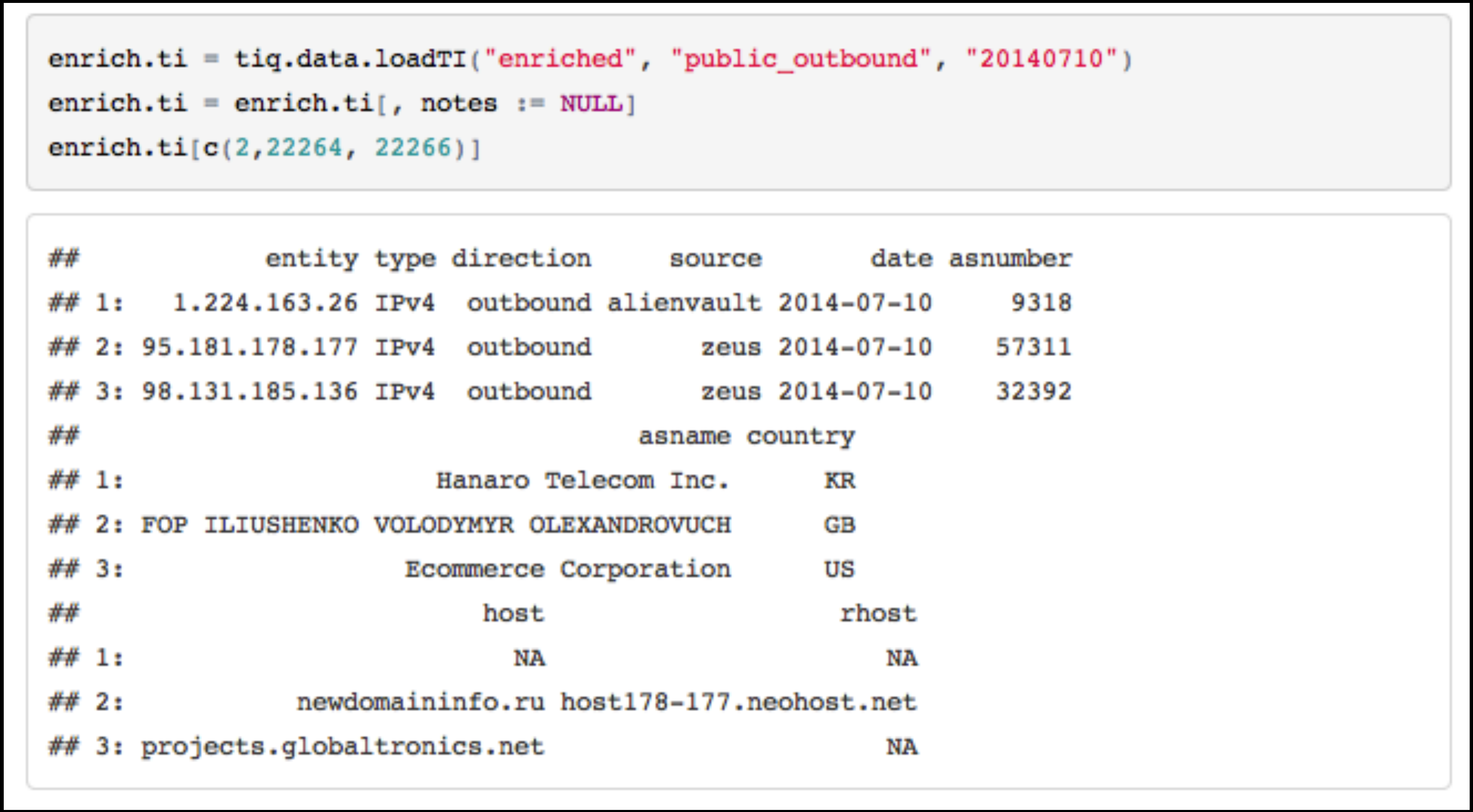
- The experiments will be around ASNs and Geolocation

Data Preparation and Cleansing

- However, we will NOT be using maps. Just let it go.

Data Preparation and Cleansing

- Small enriched sample:



```
enrich.ti = tiq.data.loadTI("enriched", "public_outbound", "20140710")
enrich.ti = enrich.ti[, notes := NULL]
enrich.ti[c(2,22264, 22266)]
```

```
##             entity type direction     source       date asnumber
## 1:   1.224.163.26 IPv4  outbound alienvault 2014-07-10     9318
## 2: 95.181.178.177 IPv4  outbound       zeus 2014-07-10    57311
## 3: 98.131.185.136 IPv4  outbound       zeus 2014-07-10    32392
##                                     asname country
## 1:                     Hanaro Telecom Inc.      KR
## 2: FOP ILIUSHENKO VOLODYMYR OLEXANDROVUCH      GB
## 3:                   Ecommerce Corporation      US
##                           host                   rhost
## 1:                         NA                      NA
## 2:           newdomaininfo.ru host178-177.neohost.net
## 3: projects.globaltronics.net                      NA
```

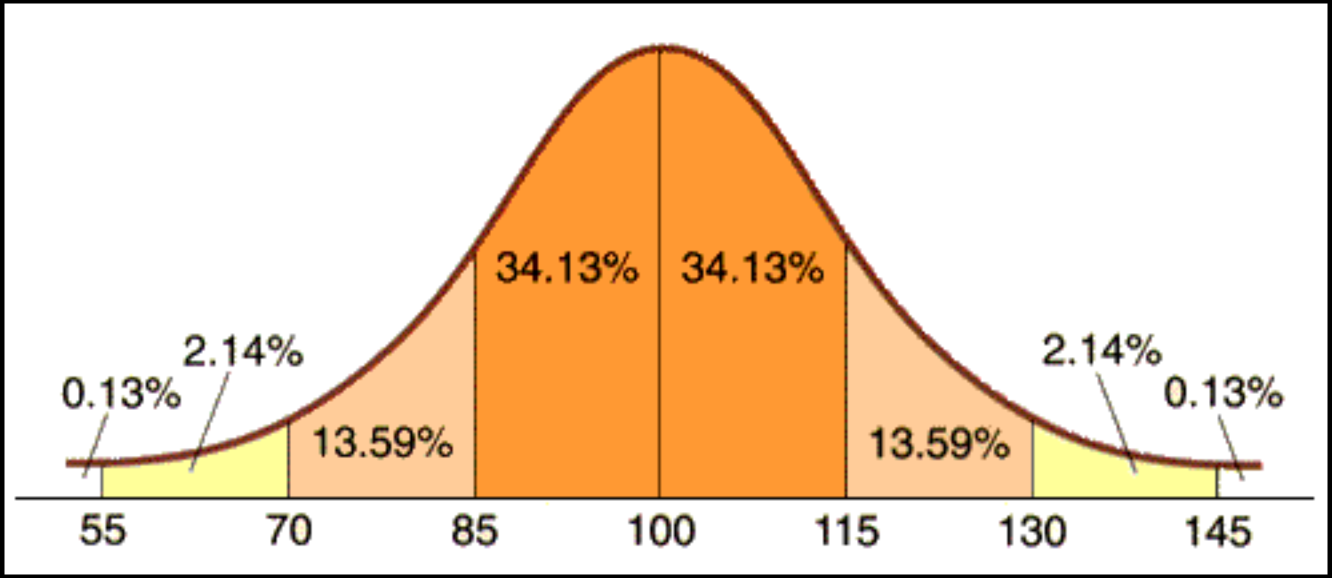
Testing the Data

- Let’s generate some interesting metrics:
    - NOVELTY – How often do they update themselves?
    - OVERLAP – How do they compare to what you got?
    - POPULATION – what is in them anyway?

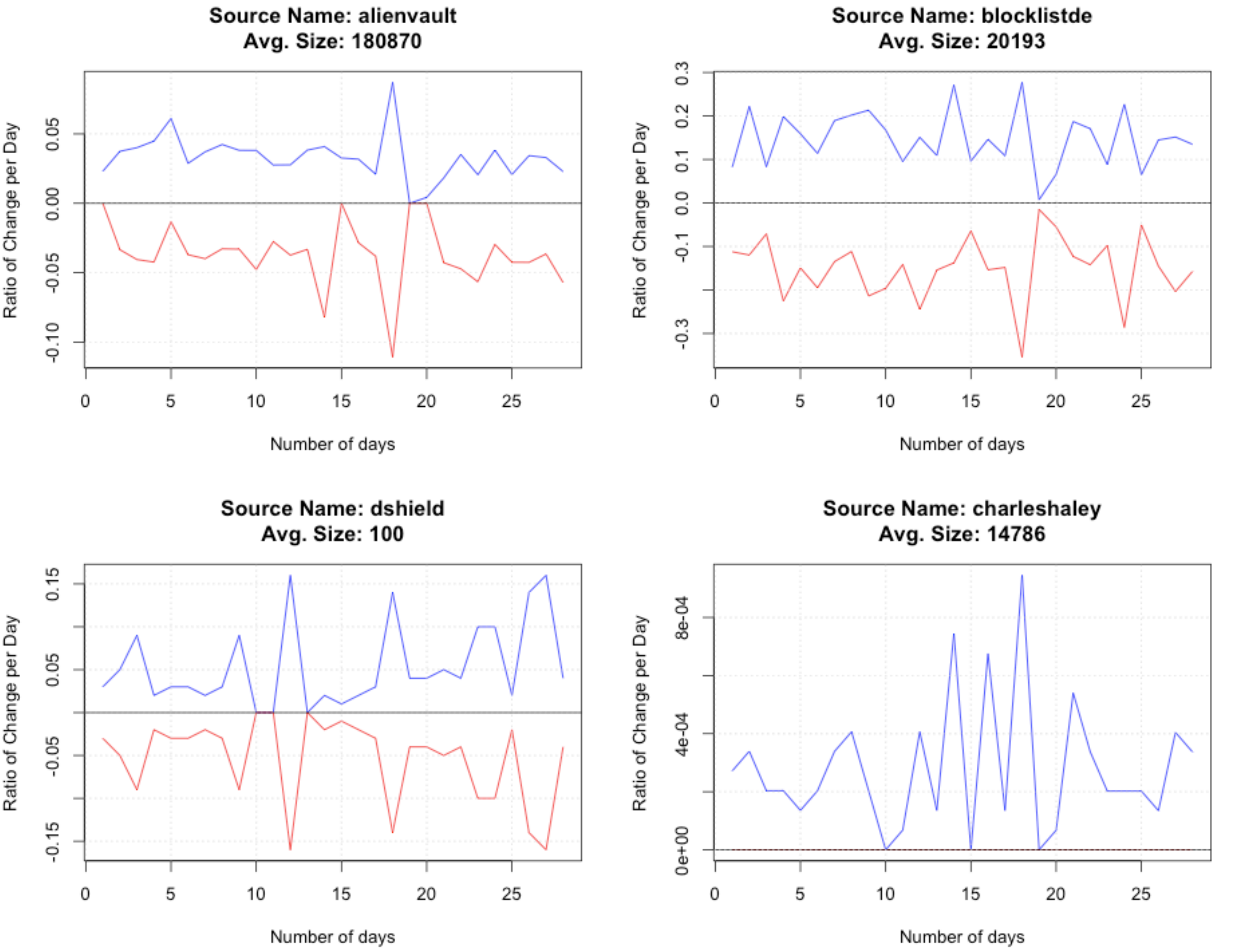
- Population is tricky:
    - Could mean the entire world (all IPv4 space)
    - Should ideally mean YOUR world

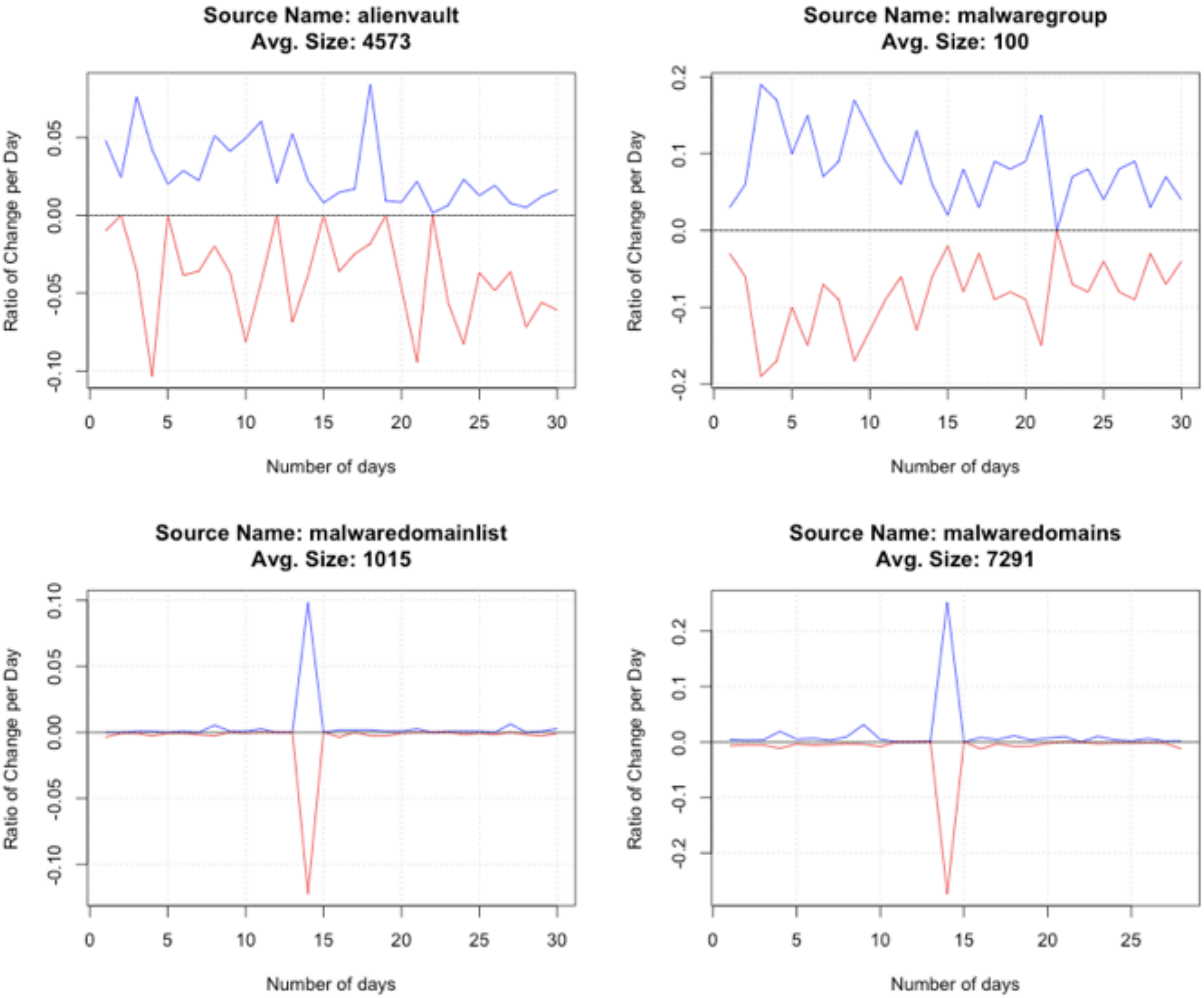
But WHAT IS THE IQ?!??1?

- We will withhold judgment
- The best data composition is the best one for you
- We will do our best to explain results so you can decide.
- Maybe on further (or more private) research…

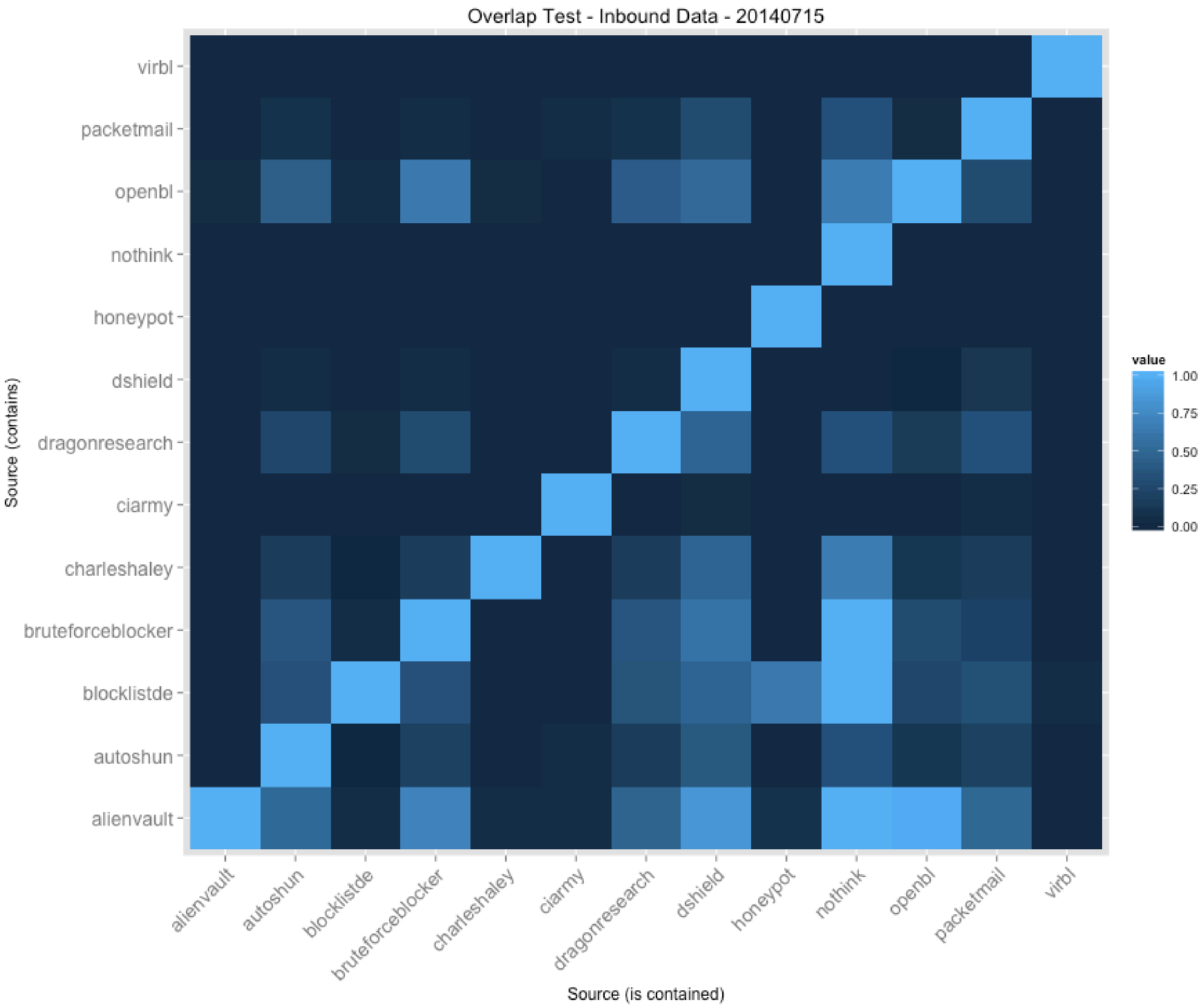


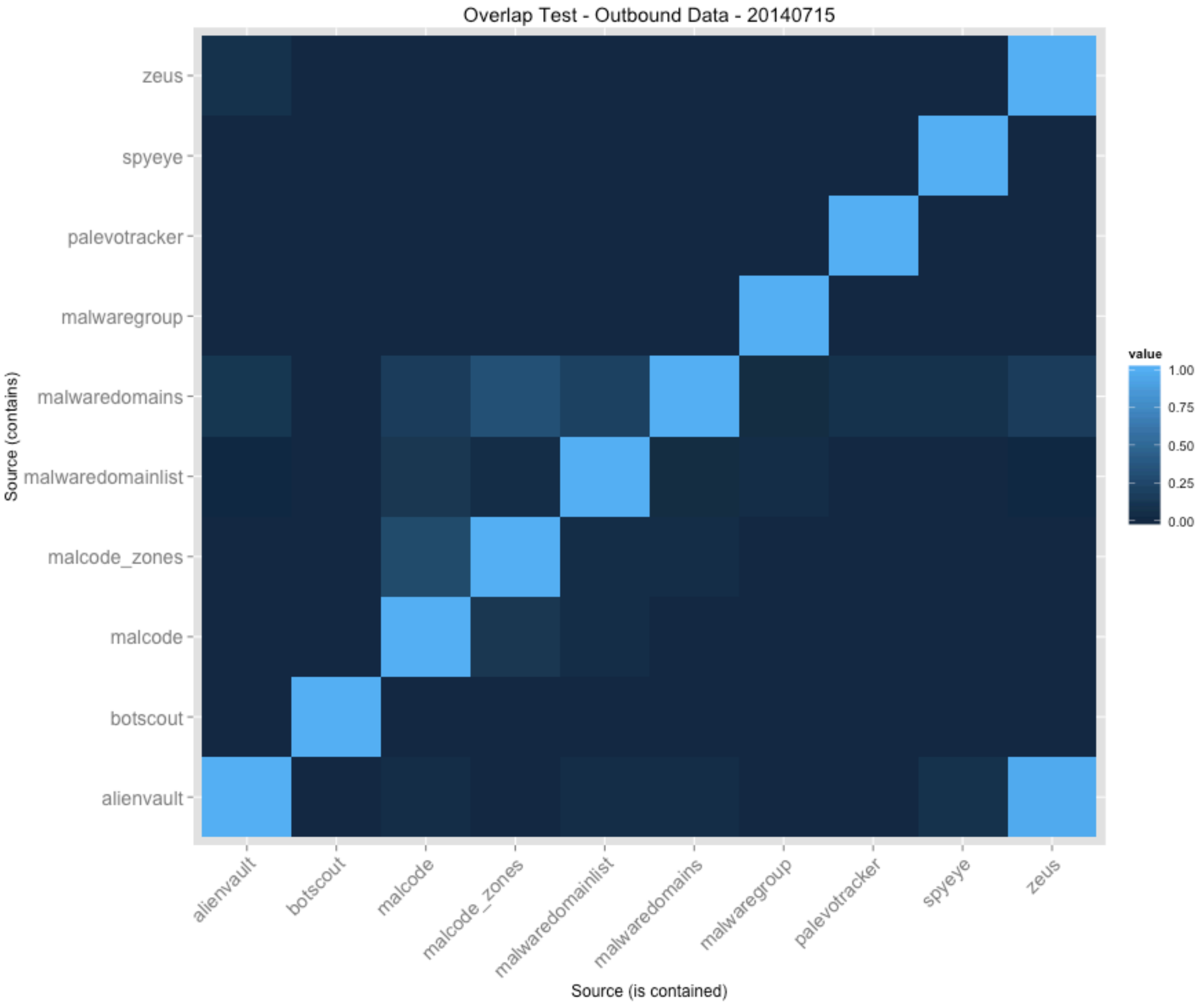
Novelty Test – measuring added and dropped indicators

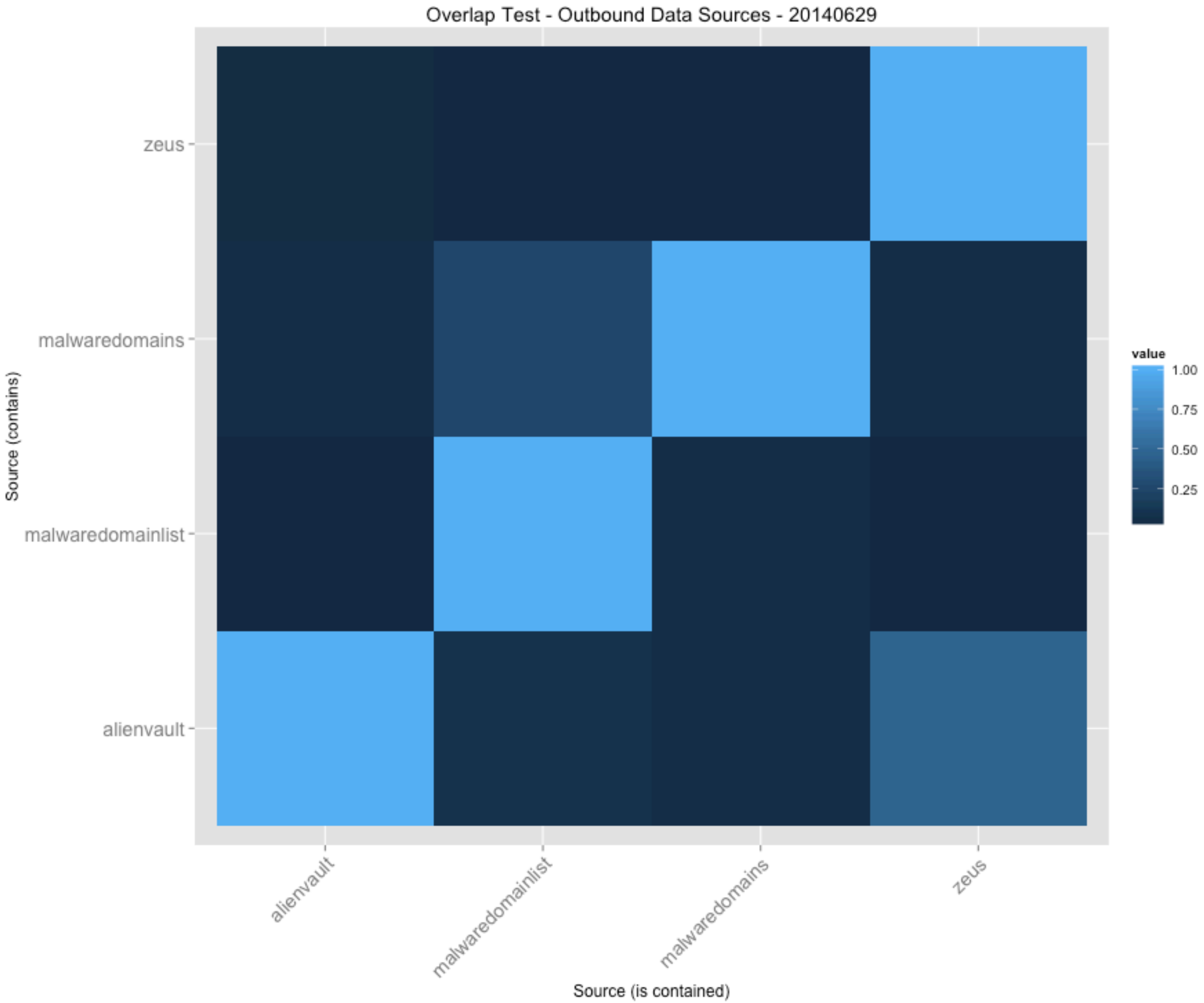

Source Name: alienvault
Avg. Size: 180870
Ratio of Change per Day
-0.10
-0.05
0.00
0.05
0
5
10
15
20
25
Number of days
Source Name: blocklistde
Avg. Size: 20193
Ratio of Change per Day
-0.3
-0.1
0.0
0.1
0.2
0.3
0
5
10
15
20
25
Number of days
Source Name: dshield
Avg. Size: 100
Ratio of Change per Day
-0.15
-0.05
0.05
0.15
0
5
10
15
20
25
Number of days
Source Name: charleshaley
Avg. Size: 14786
Ratio of Change per Day
0e+00
4e-04
8e-04
0
5
10
15
20
25
Number of days


Source Name: alienvault
Avg. Size: 4573
Ratio of Change per Day
0.05
0.00
-0.05
-0.10
0
5
10
15
20
25
30
Number of days
Source Name: malwaregroup
Avg. Size: 100
Ratio of Change per Day
0.2
0.1
0.0
-0.1
-0.2
0
5
10
15
20
25
30
Number of days
Source Name: malwaredomainlist
Avg. Size: 1015
Ratio of Change per Day
0.10
0.05
0.00
-0.05
-0.10
0
5
10
15
20
25
30
Number of days
Source Name: malwaredomains
Avg. Size: 7291
Ratio of Change per Day
0.2
0.1
0.0
-0.1
-0.2
0
5
10
15
20
25
Number of days

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


Overlap Test - Inbound Data - 20140715
Source (contains)
virbl
packetmail
openbl
nothink
honeypot
dshield
dragonresearch
ciarmy
charleshaley
bruteforceblocker
blocklistde
autoshun
alienvault
alienvault
autoshun
blocklistde
bruteforceblocker
charleshaley
ciarmy
dragonresearch
dshield
honeypot
nothink
openbl
packetmail
virbl
Source (is contained)
value
1.00
0.75
0.50
0.25
0.00

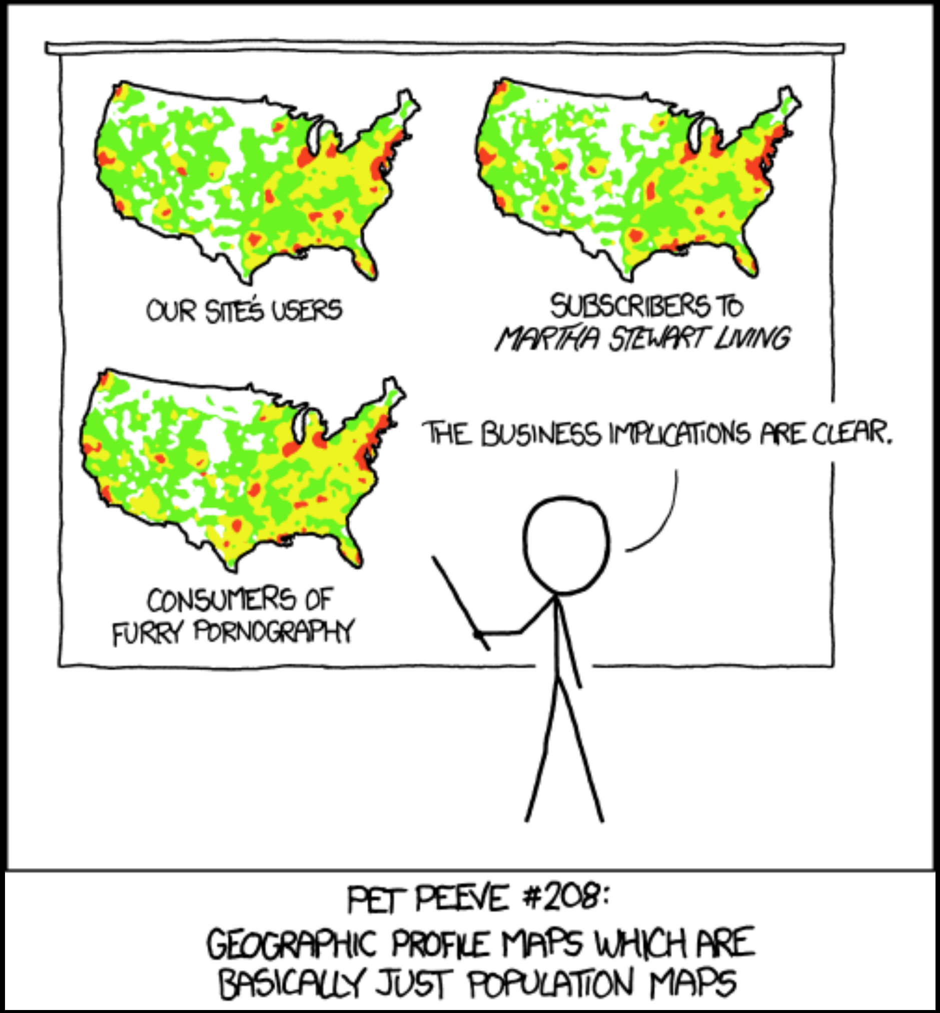

Overlap Test - Outbound Data - 20140715
Source (contains)
zeus
spyeye
palevotracker
malwaregroup
malwaredomains
malwaredomainlist
malcode_zones
malcode
botscout
alienvault
Source (is contained)
alienvault
botscout
malcode
malcode_zones
malwaredomainlist
malwaredomains
malwaregroup
palevotracker
spyeye
zeus
value
1.00
0.75
0.50
0.25
0.00

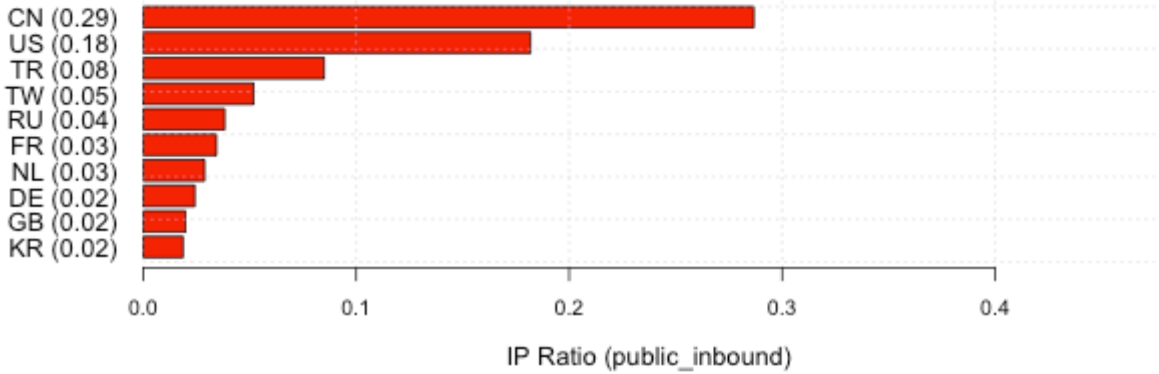

Overlap Test - Outbound Data Sources - 20140629
Source (contains)
zeus
malwaredomains
malwaredomainlist
alienvault
Source (is contained)
alienvault
malwaredomainlist
malwaredomains
zeus
value
1.00
0.75
0.50
0.25

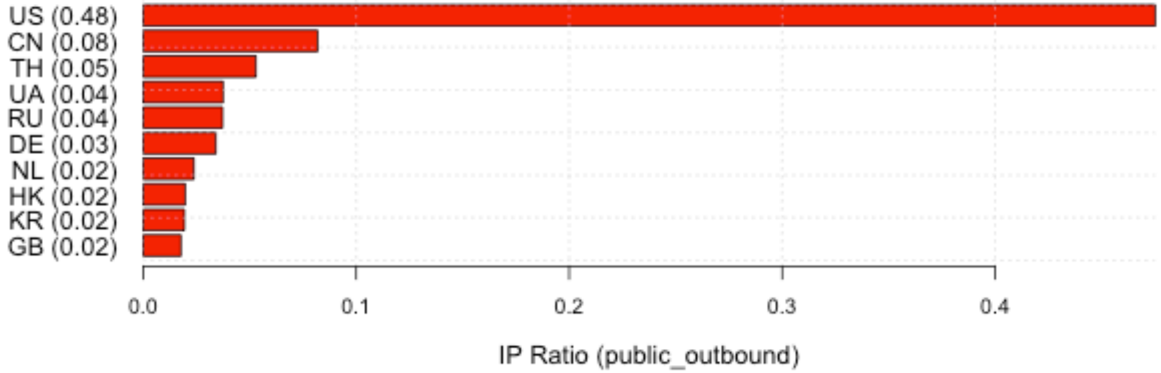
Population Test

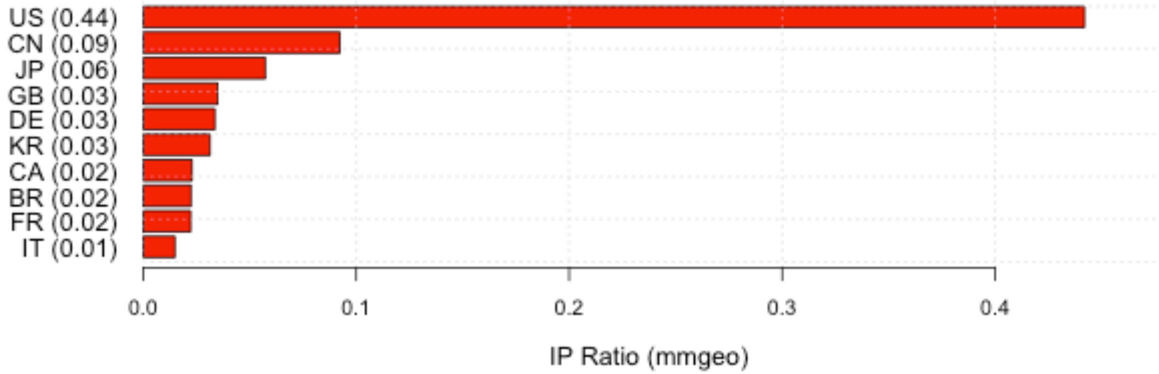
- Let us use the ASN and GeoIP databases that we used to enrich our data as a reference of the “true” population.

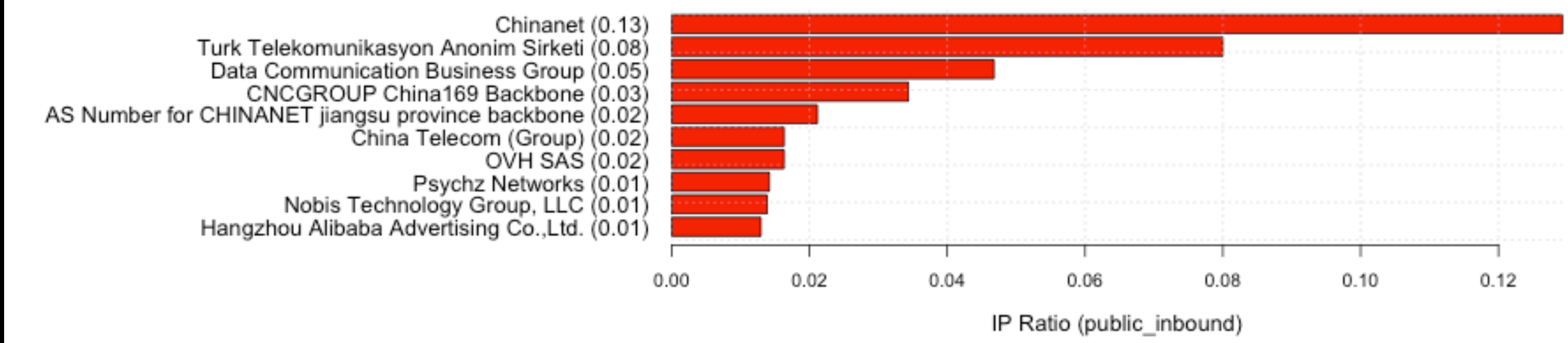
- But, but, human beings are unpredictable! We will never be able to forecast this!

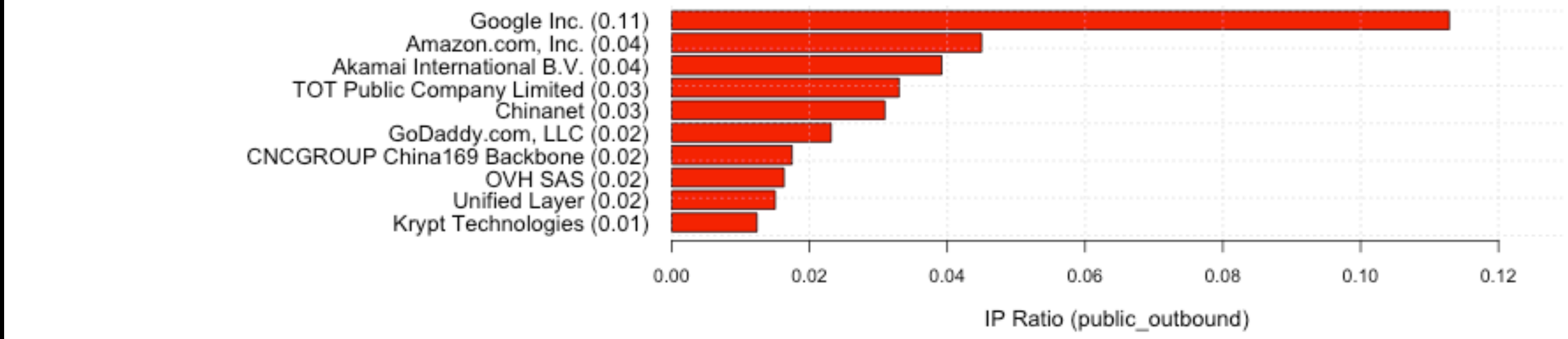

OUR SITE'S USERS
SUBSCRIBERS TO MARTHA STEWART LIVING
CONSUMERS OF FURRY PORNOGRAPHY
THE BUSINESS IMPLICATIONS ARE CLEAR.
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

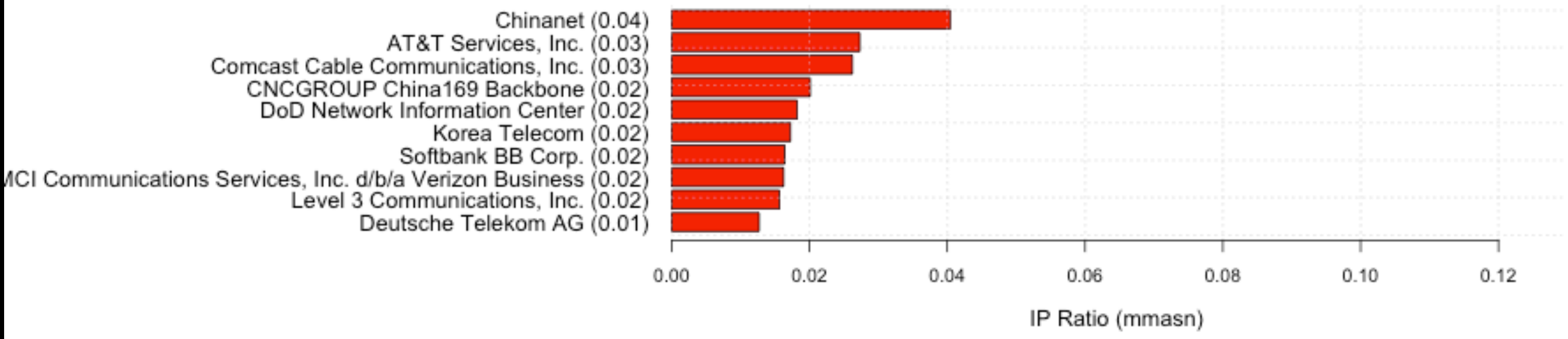

CN (0.29)
US (0.18)
TR (0.08)
TW (0.05)
RU (0.04)
FR (0.03)
NL (0.03)
DE (0.02)
GB (0.02)
KR (0.02)
0.0
0.1
0.2
0.3
0.4
IP Ratio (public_inbound)


US (0.48)
CN (0.08)
TH (0.05)
UA (0.04)
RU (0.04)
DE (0.03)
NL (0.02)
HK (0.02)
KR (0.02)
GB (0.02)
0.0
0.1
0.2
0.3
0.4
IP Ratio (public_outbound)


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 (mmgeo)


Chinanet (0.13)
Turk Telekomunikasyon Anonim Sirketi (0.08)
Data Communication Business Group (0.05)
CNCGROUP China169 Backbone (0.03)
AS Number for CHINANET jiangsu province backbone (0.02)
China Telecom (Group) (0.02)
OVH SAS (0.02)
Psychz Networks (0.01)
Nobis Technology Group, LLC (0.01)
Hangzhou Alibaba Advertising Co.,Ltd. (0.01)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
IP Ratio (public_inbound)


Google Inc. (0.11)
Amazon.com, Inc. (0.04)
Akamai International B.V. (0.04)
TOT Public Company Limited (0.03)
Chinanet (0.03)
GoDaddy.com, LLC (0.02)
CNCGROUP China169 Backbone (0.02)
OVH SAS (0.02)
Unified Layer (0.02)
Krypt Technologies (0.01)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
IP Ratio (public_outbound)

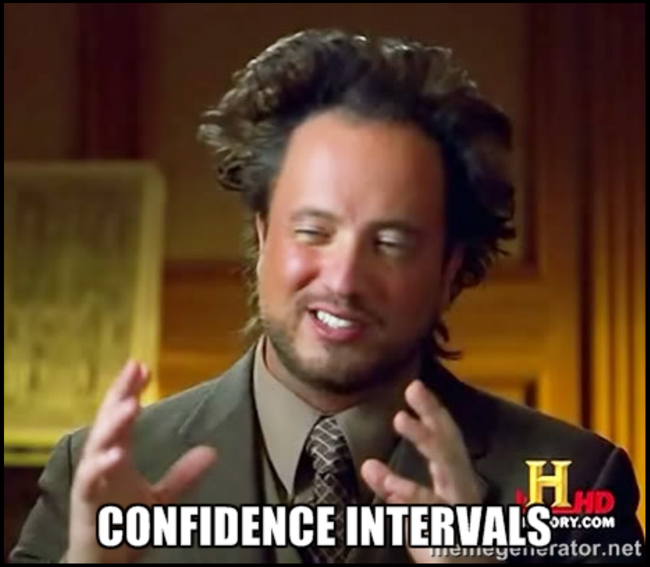

Chinanet (0.04)
AT&T Services, Inc. (0.03)
Comcast Cable Communications, Inc. (0.03)
CNCGROUP China169 Backbone (0.02)
DoD Network Information Center (0.02)
Korea Telecom (0.02)
Softbank BB Corp. (0.02)
MCI Communications Services, Inc. d/b/a Verizon Business (0.02)
Level 3 Communications, Inc. (0.02)
Deutsche Telekom AG (0.01)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
IP Ratio (mmasn)

Can we get a better look?

- Don't like squinting either
- Statistical inference-based comparison models (hypothesis testing)
  - Exact binomial tests (when we have the "true" pop)
  - Chi-squared proportion tests (similar to independence tests)

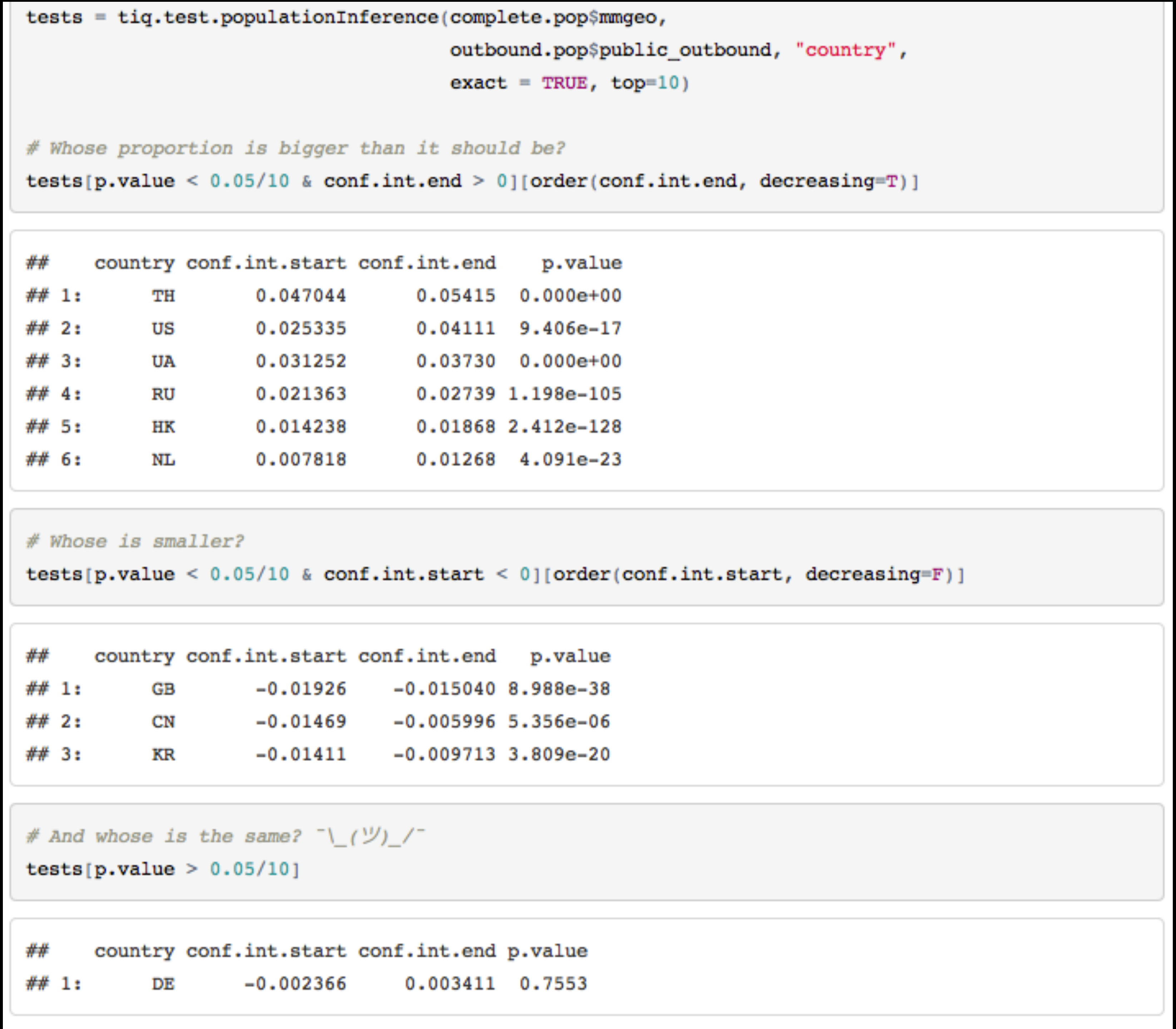
Can we get a better look?

- We can better estimate, with confidence intervals, our measures of error.
- Also, p-values! (with apologies to Alex Hutton)
- We promise to be very conservative in using them.



Doctor's discovery revealed the secret to learning any problem with just 10 training samples. Watch this shocking video and learn how rapidly you can find a solution to your learning problems using this one sneaky kernel trick! Free from overfitting!

http://www.oneweirdkerneltrick.com



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

# Whose proportion is bigger than it should be?
tests[p.value < 0.05/10 & conf.int.end > 0][order(conf.int.end, decreasing=T)]
```

```
##    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
## 3:      UA       0.031252      0.03730  0.000e+00
## 4:      RU       0.021363      0.02739 1.198e-105
## 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
## 3:      KR       -0.01411    -0.009713 3.809e-20
```

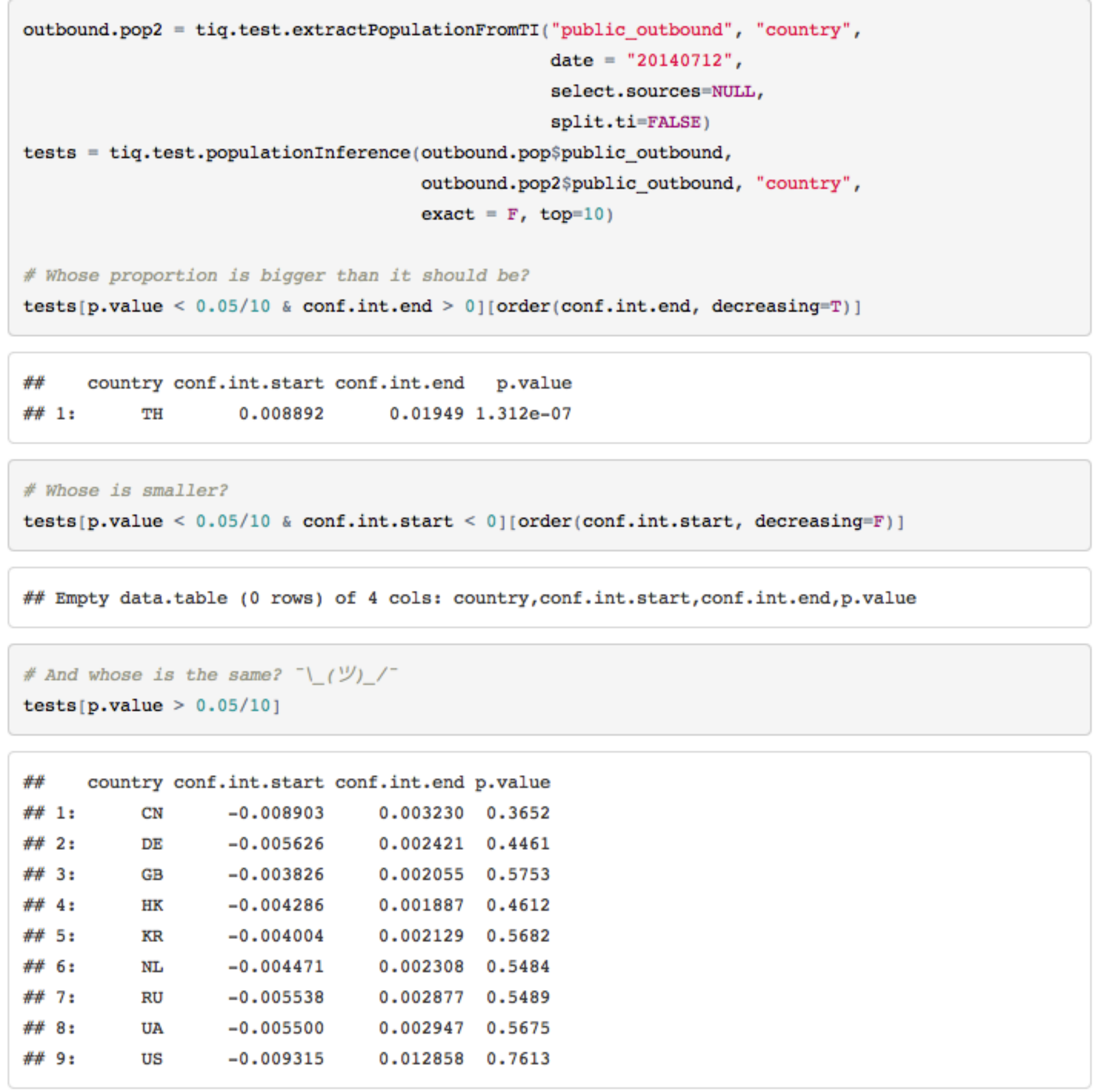
```
# And whose is the same? ¯\_(ツ)_/¯
tests[p.value > 0.05/10]
```

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

Hacker Spirit Animal™ Guide

- US – Eagle
- CA – Moose
- FR – Frog
- GB – Bulldog
- AU – Koala
- BR – Capybara / Toucan
- Texas – Armadillo

- Disclaimer: we do not endorse Geolocation-based attribution



```
outbound.pop2 = tiq.test.extractPopulationFromTI("public_outbound", "country",
                                                 date = "20140712",
                                                 select.sources=NULL,
                                                 split.ti=FALSE)
tests = tiq.test.populationInference(outbound.pop$public_outbound,
                                     outbound.pop2$public_outbound, "country",
                                     exact = F, top=10)

# Whose proportion is bigger than it should be?
tests[p.value < 0.05/10 & conf.int.end > 0][order(conf.int.end, decreasing=T)]
```

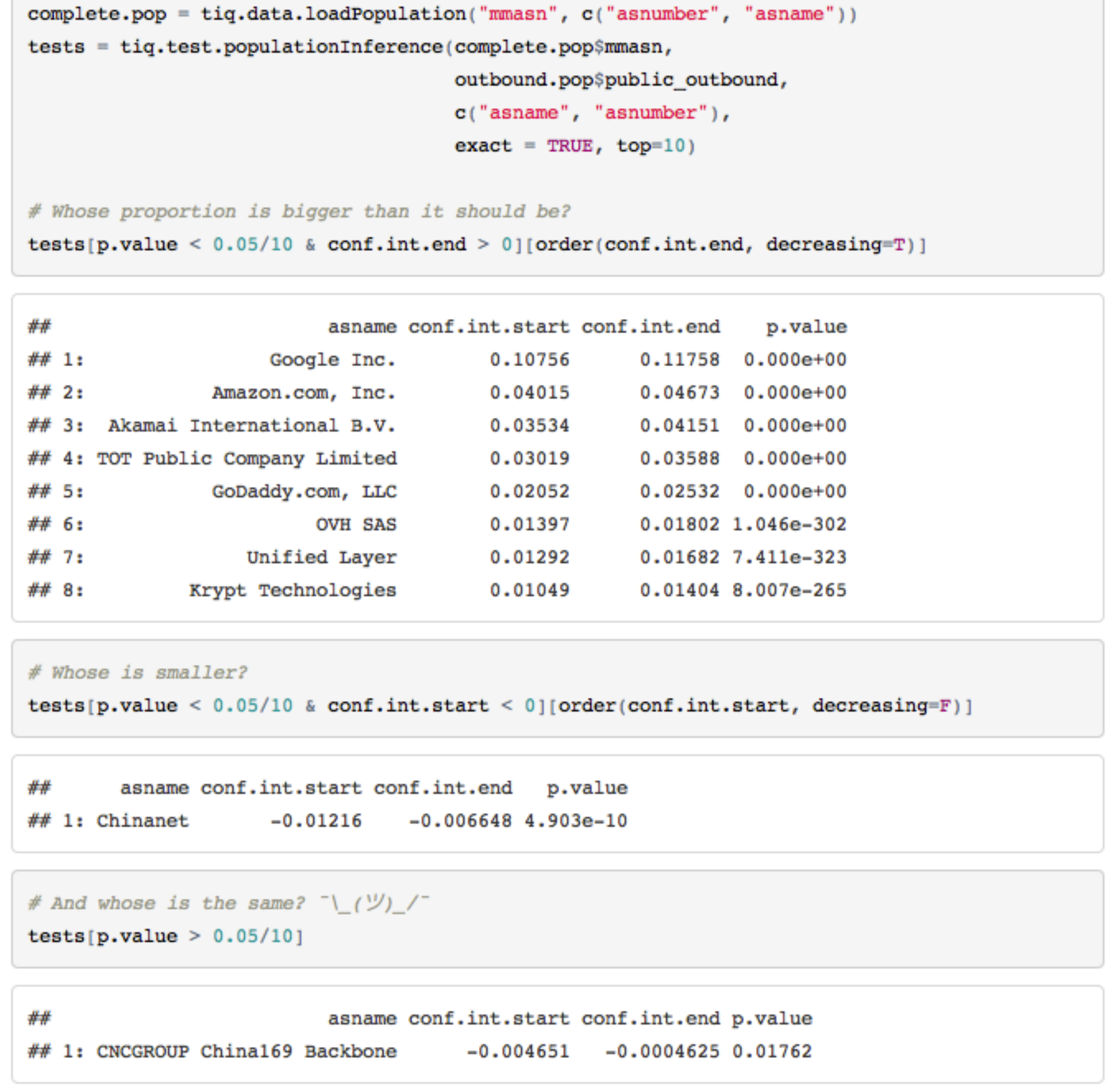
```
##    country conf.int.start conf.int.end   p.value
## 1:      TH       0.008892      0.01949 1.312e-07
```

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

```
## Empty data.table (0 rows) of 4 cols: country,conf.int.start,conf.int.end,p.value
```

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

```
##    country conf.int.start conf.int.end p.value
## 1:      CN      -0.008903     0.003230  0.3652
## 2:      DE      -0.005626     0.002421  0.4461
## 3:      GB      -0.003826     0.002055  0.5753
## 4:      HK      -0.004286     0.001887  0.4612
## 5:      KR      -0.004004     0.002129  0.5682
## 6:      NL      -0.004471     0.002308  0.5484
## 7:      RU      -0.005538     0.002877  0.5489
## 8:      UA      -0.005500     0.002947  0.5675
## 9:      US      -0.009315     0.012858  0.7613
```



```
complete.pop = tiq.data.loadPopulation("mmasn", c("asnumber", "asname"))
tests = tiq.test.populationInference(complete.pop$mmasn,
                                     outbound.pop$public_outbound,
                                     c("asname", "asnumber"),
                                     exact = TRUE, top=10)

# Whose proportion is bigger than it should be?
tests[p.value < 0.05/10 & conf.int.end > 0][order(conf.int.end, decreasing=T)]
```

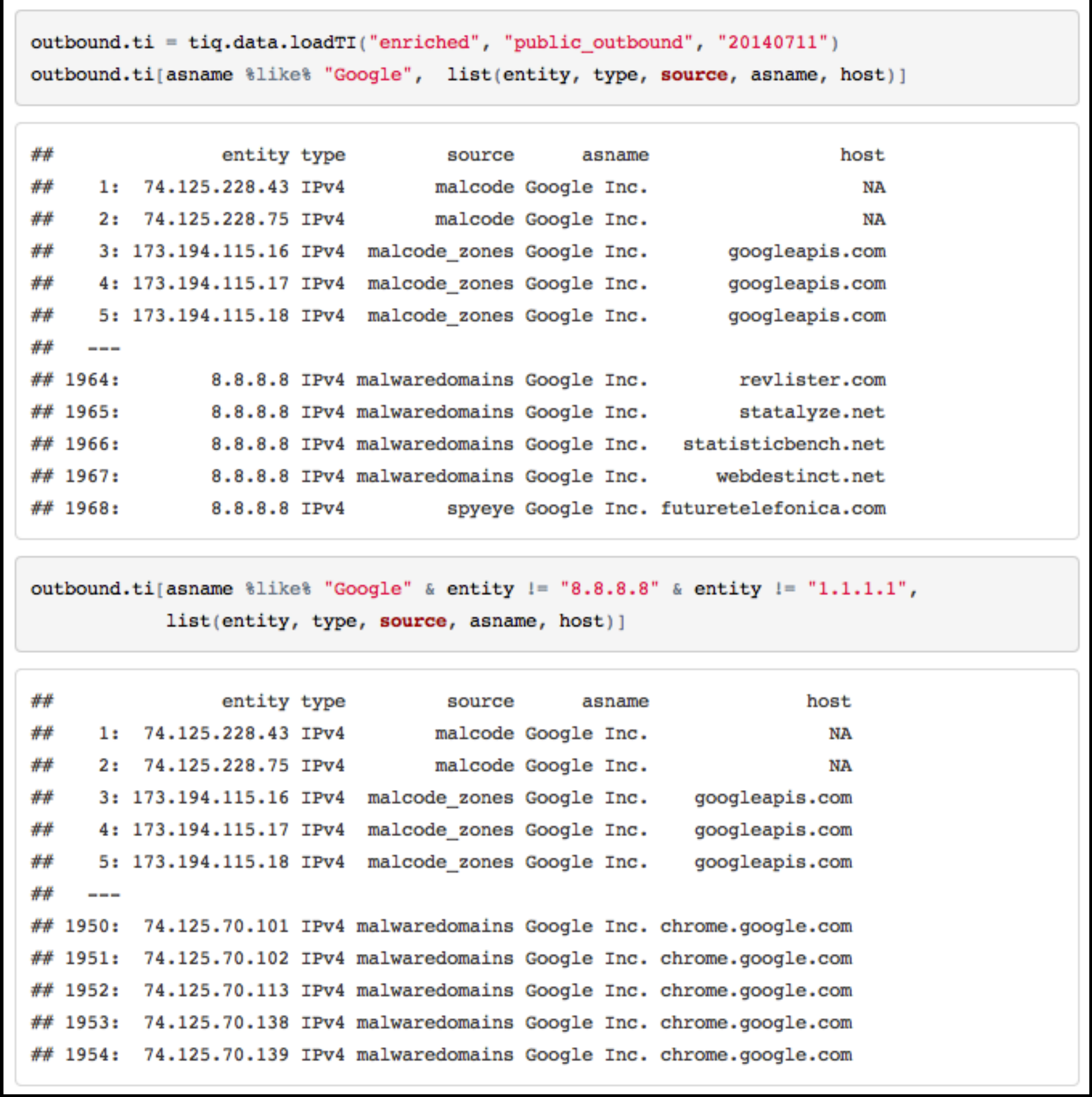
```
##                        asname conf.int.start conf.int.end    p.value
## 1:                Google Inc.        0.10756      0.11758  0.000e+00
## 2:           Amazon.com, Inc.        0.04015      0.04673  0.000e+00
## 3:  Akamai International B.V.        0.03534      0.04151  0.000e+00
## 4: TOT Public Company Limited        0.03019      0.03588  0.000e+00
## 5:           GoDaddy.com, LLC        0.02052      0.02532  0.000e+00
## 6:                    OVH SAS        0.01397      0.01802 1.046e-302
## 7:              Unified Layer        0.01292      0.01682 7.411e-323
## 8:        Krypt Technologies        0.01049      0.01404 8.007e-265
```

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

```
##      asname conf.int.start conf.int.end   p.value
## 1: Chinanet       -0.01216    -0.006648 4.903e-10
```

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

```
##                        asname conf.int.start conf.int.end p.value
## 1: CNCGROUP China169 Backbone      -0.004651   -0.0004625 0.01762
```



```
outbound.ti = tiq.data.loadTI("enriched", "public_outbound", "20140711")
outbound.ti[asname %like% "Google",  list(entity, type, source, asname, host)]
```

```
##               entity type         source      asname                 host
##    1:  74.125.228.43 IPv4        malcode Google Inc.                   NA
##    2:  74.125.228.75 IPv4        malcode Google Inc.                   NA
##    3: 173.194.115.16 IPv4  malcode_zones Google Inc.       googleapis.com
##    4: 173.194.115.17 IPv4  malcode_zones Google Inc.       googleapis.com
##    5: 173.194.115.18 IPv4  malcode_zones Google Inc.       googleapis.com
##   ---
## 1964:        8.8.8.8 IPv4 malwaredomains Google Inc.        revlister.com
## 1965:        8.8.8.8 IPv4 malwaredomains Google Inc.        statalyze.net
## 1966:        8.8.8.8 IPv4 malwaredomains Google Inc.   statisticbench.net
## 1967:        8.8.8.8 IPv4 malwaredomains Google Inc.      webdestinct.net
## 1968:        8.8.8.8 IPv4         spyeye Google Inc. futuretelefonica.com
```

```
outbound.ti[asname %like% "Google" & entity != "8.8.8.8" & entity != "1.1.1.1",
            list(entity, type, source, asname, host)]
```

```
##               entity type         source      asname              host
##    1:  74.125.228.43 IPv4        malcode Google Inc.                NA
##    2:  74.125.228.75 IPv4        malcode Google Inc.                NA
##    3: 173.194.115.16 IPv4  malcode_zones Google Inc.    googleapis.com
##    4: 173.194.115.17 IPv4  malcode_zones Google Inc.    googleapis.com
##    5: 173.194.115.18 IPv4  malcode_zones Google Inc.    googleapis.com
##   ---
## 1950:  74.125.70.101 IPv4 malwaredomains Google Inc. chrome.google.com
## 1951:  74.125.70.102 IPv4 malwaredomains Google Inc. chrome.google.com
## 1952:  74.125.70.113 IPv4 malwaredomains Google Inc. chrome.google.com
## 1953:  74.125.70.138 IPv4 malwaredomains Google Inc. chrome.google.com
## 1954:  74.125.70.139 IPv4 malwaredomains Google Inc. chrome.google.com
```

Introducing COMBINE

- Harvesting feeds takes some work.
- Most of us let somebody else do it without thinking about what it actually takes.

Introducing COMBINE

https://github.com/mlsecproject/combine

Introducing COMBINE

- Components:
  1. **Reaper** gathers the threat data directly from feeds.
  2. **Thresher** normalizes it into a simplistic data model.
  3. **Winnower** optionally performs basic validation or enrichment.
  4. **Baler** transforms the data into CybOX, CSV, JSON, and CIM. (Only CSV and JSON work right now). Could also write others fairly easily. (nudge nudge, wink wink)

Introducing COMBINE

- Always trying to feed it more. Lots of possibilities, including your own data sources.

- We clearly do NOT endorse any included feeds.

Introducing COMBINE

- Enrichments - think metadata.
  - AS, geolocation
  - DNS resolutions courtesy of Farsight DNSDB
  - Ask them for an API key to test it, tell them Alex Pinto sent you ;)

MLSec Project

- Both projects have been released as GPLv3 by MLSec Project
- Will replace the internal versions we have on the main code

- Looking for participants and data sharing agreements
- Liked TIQ-TEST? We can benchmark your private feeds using these and other techniques

- Visit **https://www.mlsecproject.org** , message @MLSecProject or just e-mail me.

MLSec

MACHINE LEARNING SECURITY

Take Aways

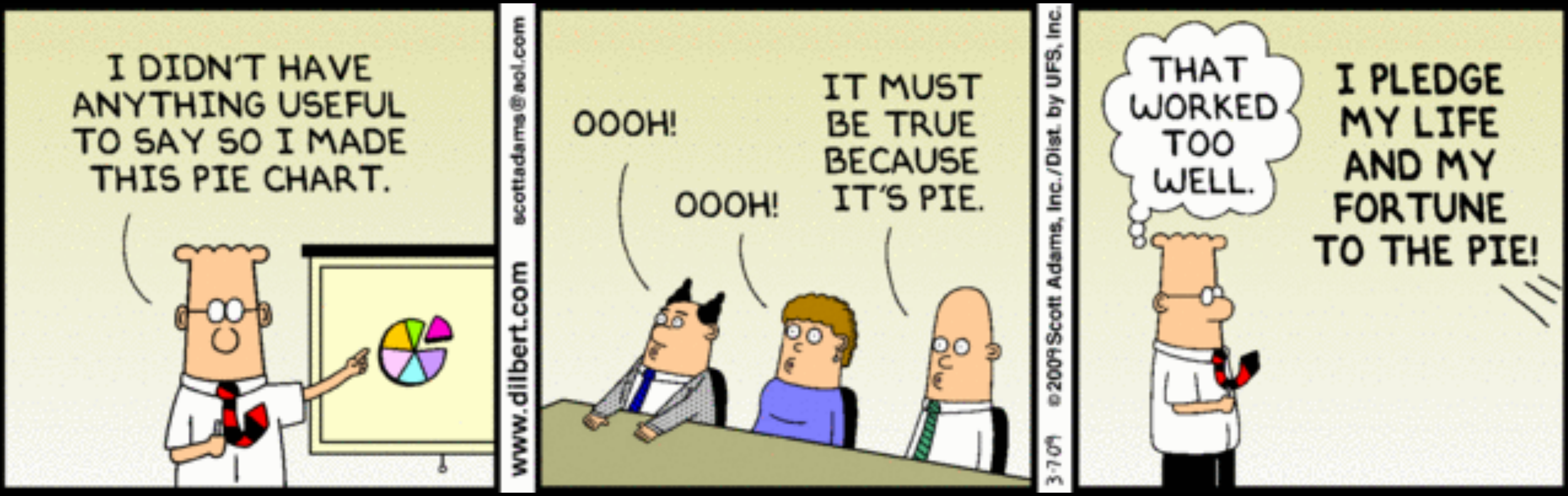
- Analyze your data.
- Extract value from it!
- Try before you buy! Different test results mean different things to different orgs.
- Use the tools! Suggest new tests!
- Share data with us! We take good care of it, make sure it gets proper exercise.

Thanks!

- Q&A?
- Feedback!

Alex Pinto
@alexcpsec
@MLSecProject

Kyle Maxwell
@kylemaxwell


I DIDN'T HAVE ANYTHING USEFUL TO SAY SO I MADE THIS PIE CHART.
OOOH!
OOOH!
IT MUST BE TRUE BECAUSE IT'S PIE.
THAT WORKED TOO WELL.
I PLEDGE MY LIFE AND MY FORTUNE TO THE PIE!
www.dilbert.com
scottadams@aol.com
3-7-09 ©2009 Scott Adams, Inc./Dist. by UFS, Inc.

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

- Albert Einstein