I, for one, welcome our new Cyber Overlords!

An introduction to the use of data science in cybersecurity

By Tiago Henriques, Filipa Rodrigues
Florentino Bexiga, Ana Barbosa

Agenda

- Who are we?
- Machine Learning and Cybersecurity
- Image Workflow
- Image Analysis in detail
- Data Visualisation

Presenter

Tiago Henriques

Tiago is the CEO and Data necromancer at BinaryEdge however he gets to meddle in the intersection of data science and cybersecurity by providing his team with lovely problems that they solve on a daily basis.

Presenter

Florentino Bexiga

Florentino is the Data MacGyver at BinaryEdge. On a daily basis he needs to deploy infrastructure used to analyse big and realtime data. When not doing that, he can be found creating models to analyse data. Give him an orange, he'll give you a skynet. Why an orange you ask? He's hungry and likes oranges, there!

Presenter

Filipa Rodrigues

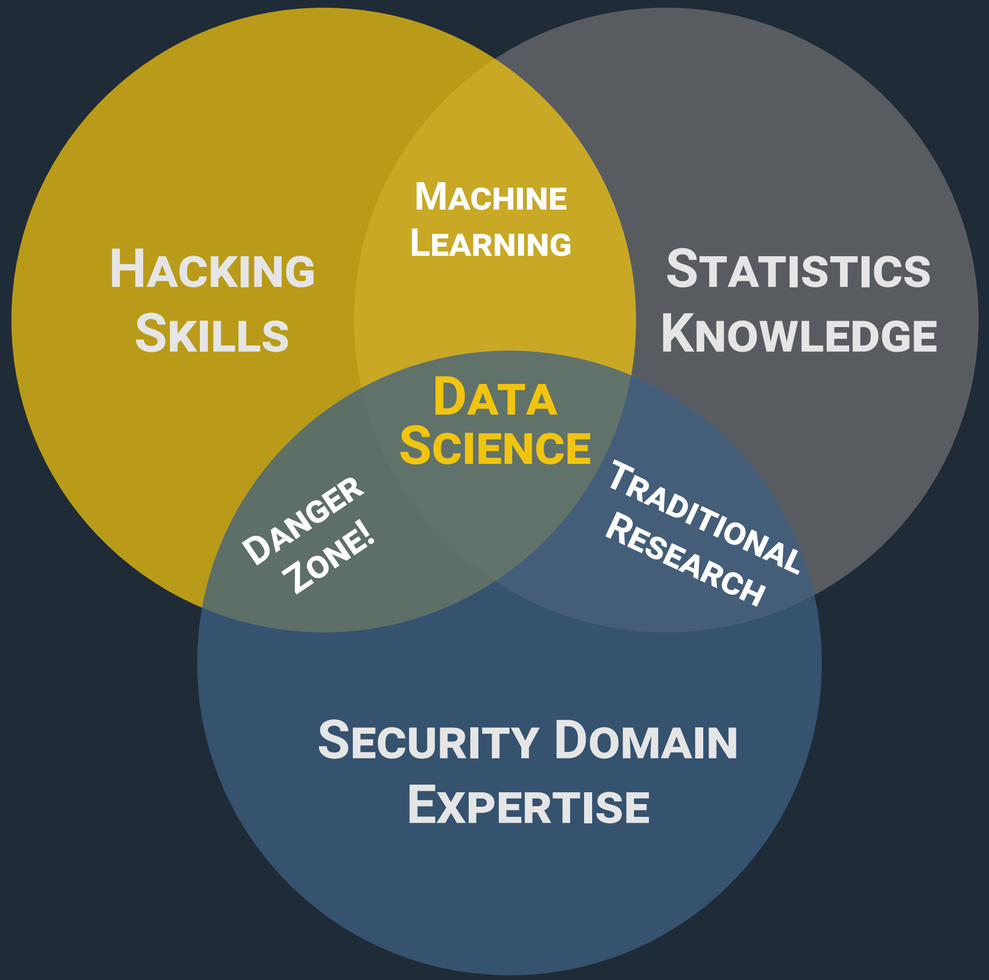
Filipa is the Data Diva at BinaryEdge, she dances the macarena with numbers to get them to tell her all their dirty secret.

Presenter

Ana Barbosa

Ana is the Data Ferret at BinaryEdge.
She is small and hides between the 110th and 111th characters of the ascii code to see and show data in that unique perspective of someone who can’t reach the box of cookies stored on top of the capitol 'I'

BinaryEdge



Source: Data-Driven Security: Analysis, visualisation and Dashboards (adapted)

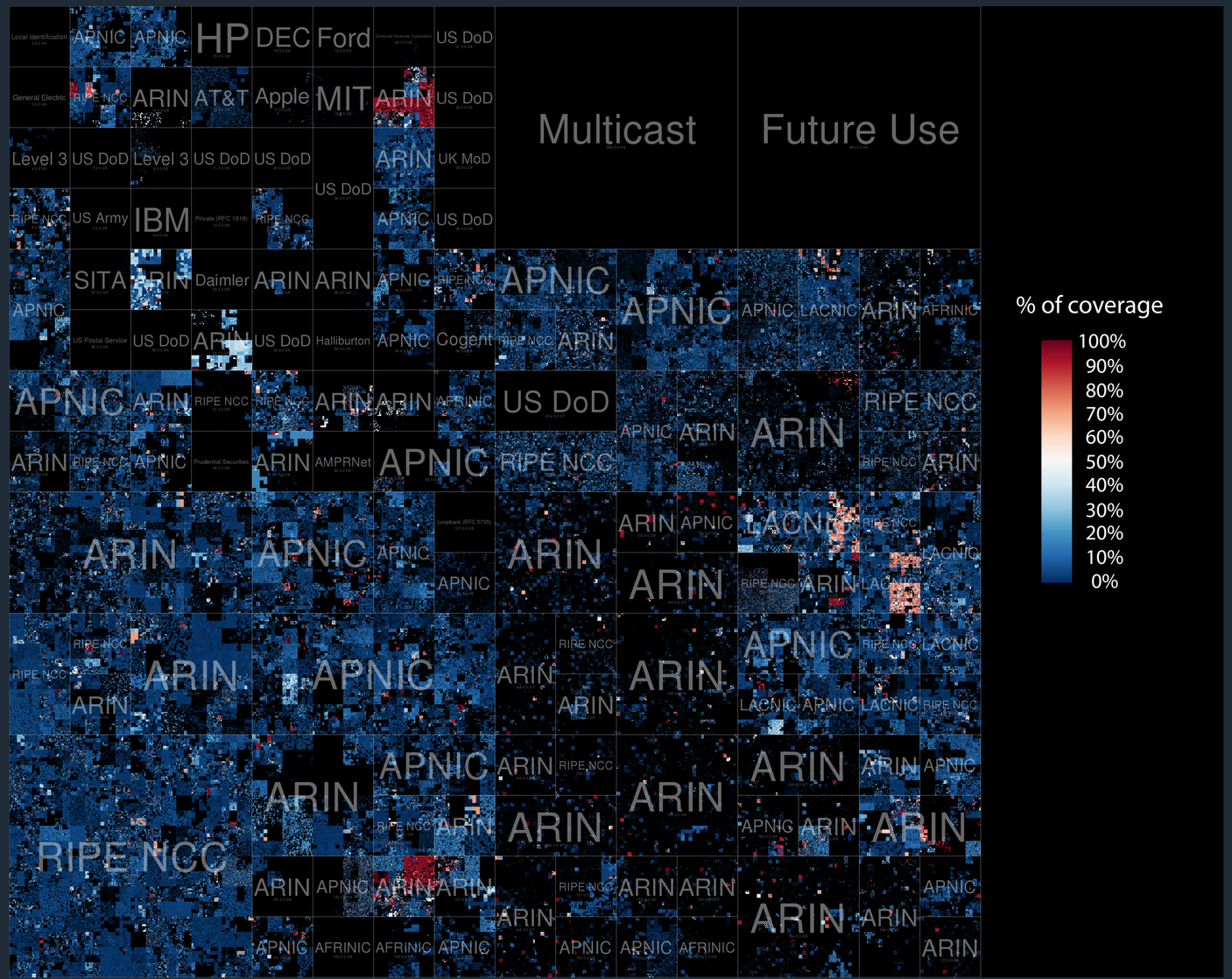
How we got here....

- **200** port scan of the entire internet/ month
- **1,400,000,000** scanning events/ month *
- **746,000** torrents monitored and increasing
- **1,362,225,600** torrent events/ month

* at a minimum

Worldwide distribution of IPs running services

Map IPv4 addresses to Hilbert curves



Data Science & Machine Learning

MULTIPLE WILD QUESTIONS APPEAR...

- How many IP addresses did job X had vs. job Y?
- What is the average duration of the scans?
- Can we extract more from all the screenshots we get?
- Can we have a more optimized job distribution?
- We can only identify X% of services because we're using static signatures, can we do better?
- Can we find similar images?

...ONE COMMON ANSWER

DATA SCIENCE

&

MACHINE LEARNING

Data Science & Machine Learning

Data Science

- Initial Analysis and Clean up
- Exploratory Data Analysis
- Data Visualisation
- Knowledge Discovery

Machine Learning

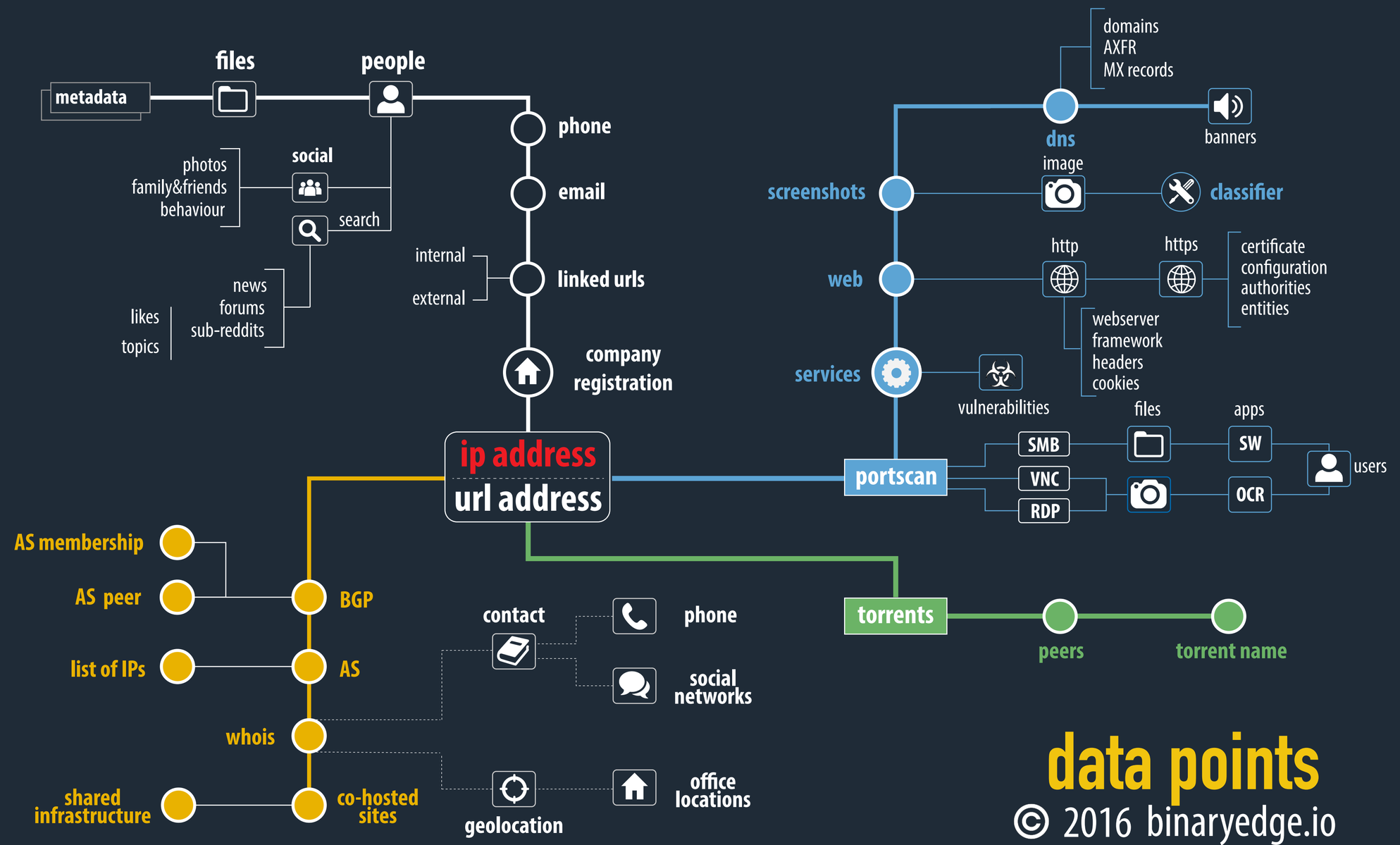
- Classification
- Clustering
- Identification
- Similarity Matching
- Regression

Problems and Limitations of Machine Learning in CyberSecurity

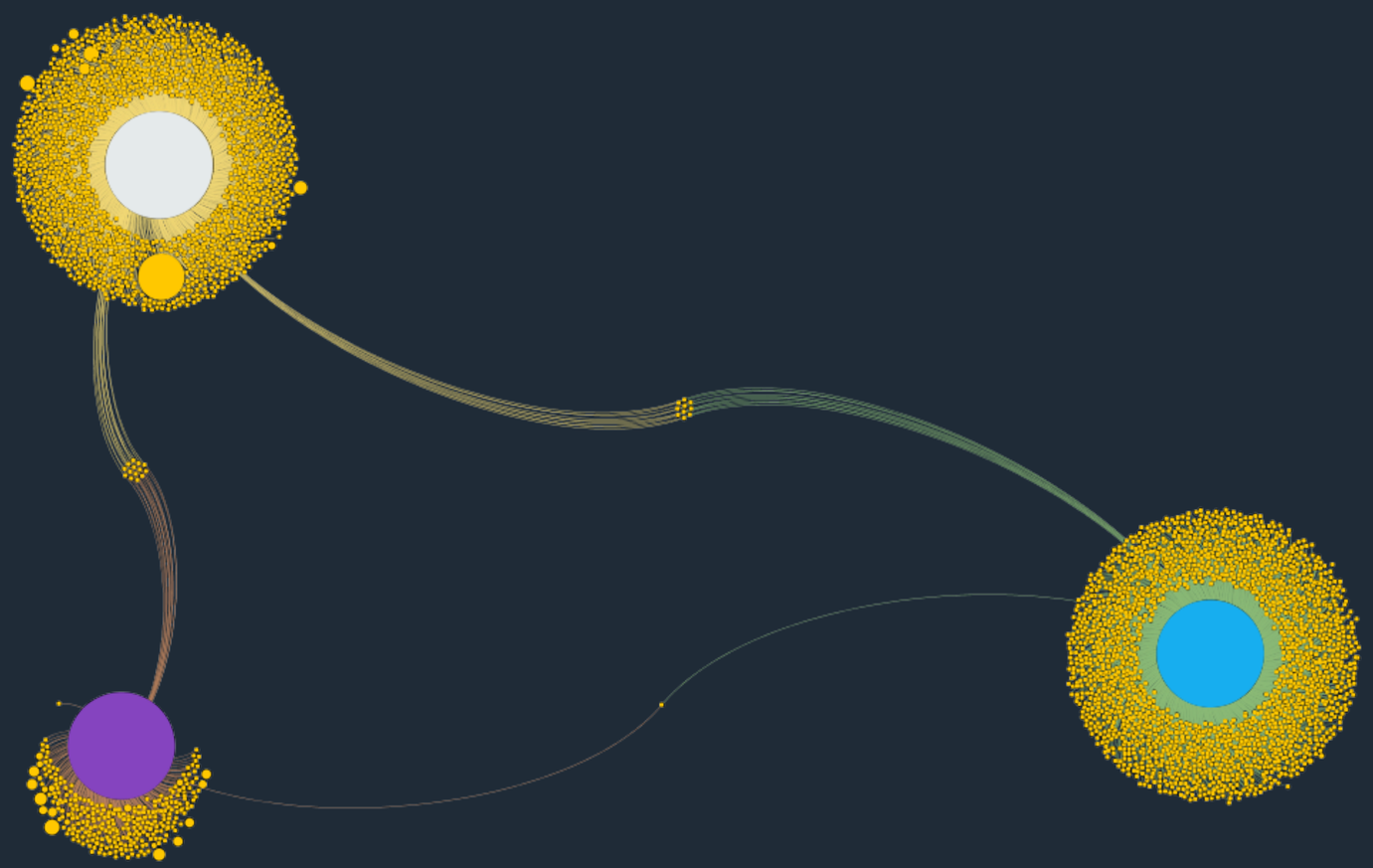
- **Lots of adversarial scenarios** – Attacks to the classifiers, goes against the foundation of machine learning
- **Prediction** – Scenarios and data too volatile, not enough proper sources of data
- **Lack of data in quantity and quality to train models**

Good use cases

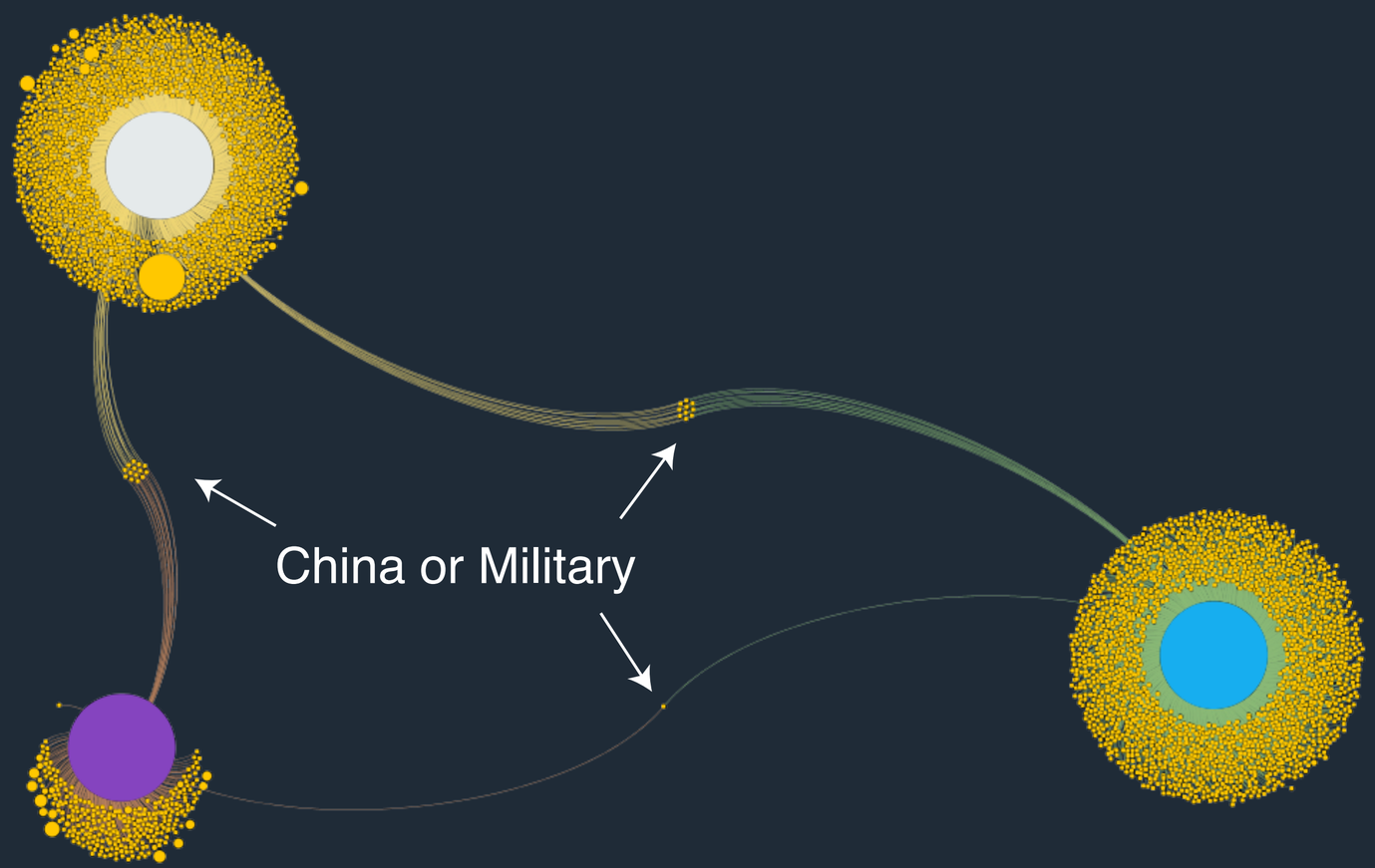
Anti-SPAM	
Antivirus	further work needs to be done, but will allow to move antivirus from a static/ signature based system into a much improved dynamic/ learning based system
Pattern detection/outlier detection (IDS/IPS)	If a computer is hacked certain behaviors will change, if constant data is being monitored and fed into a system the hack could be detected
Source code analysis	detection of vulnerable patterns during development
Smarter fuzzers	
Internal attackers	sentiment analysis applied to emails, tweets, social networks of employees


files
people
metadata
phone
email
linked urls
company registration
social
photos
family&friends
behaviour
search
news
forums
sub-reddits
likes
topics
internal
external
ip address
url address
AS membership
AS peer
BGP
list of IPs
AS
whois
shared infrastructure
co-hosted sites
contact
phone
social networks
geolocation
office locations
domains
AXFR
MX records
dns
banners
screenshots
image
classifier
web
http
https
certificate
configuration
authorities
entities
webserver
framework
headers
cookies
services
vulnerabilities
portscan
SMB
VNC
RDP
files
apps
SW
OCR
users
torrents
peers
torrent name
data points
© 2016 binaryedge.io

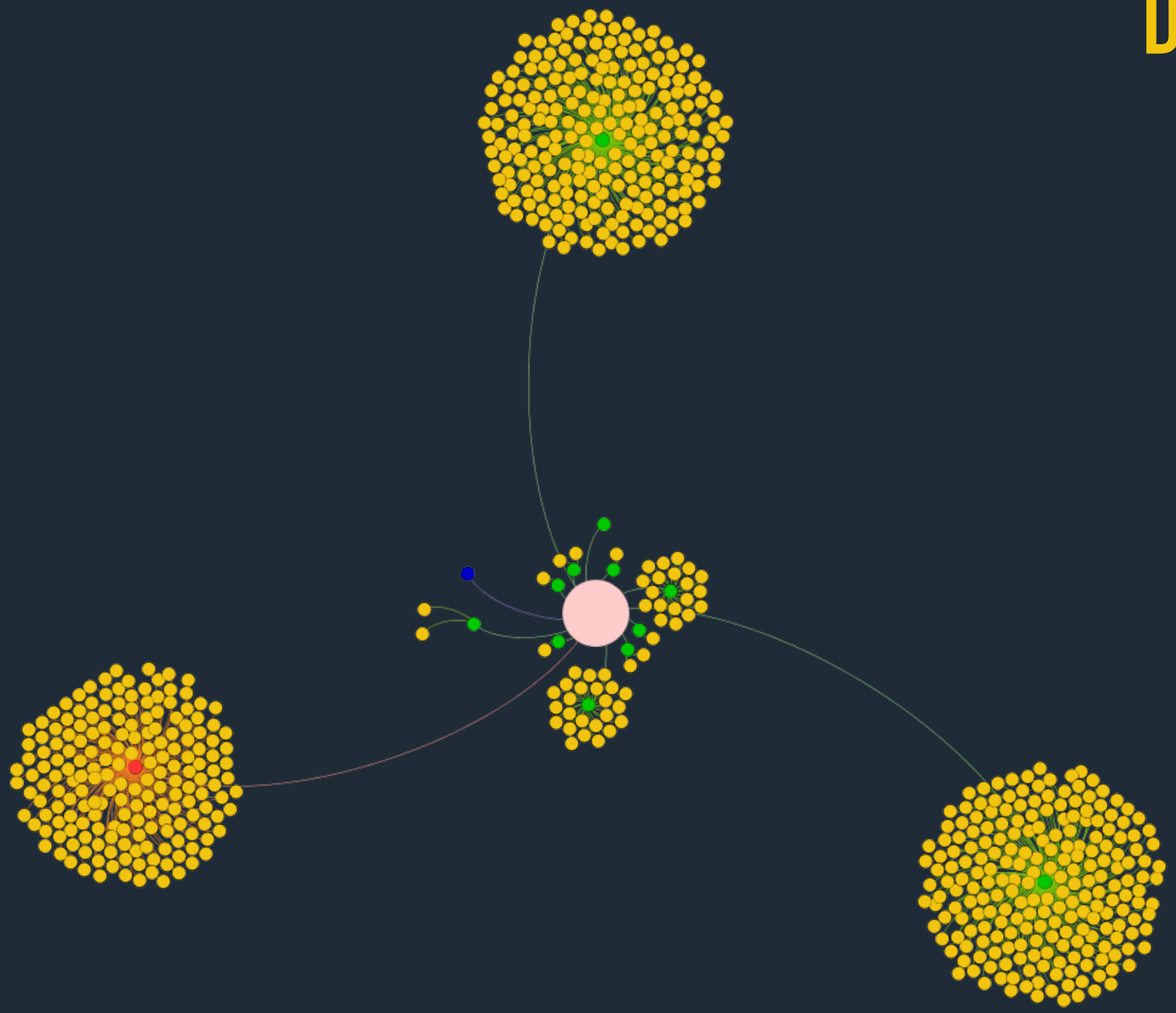
Torrent Correlation



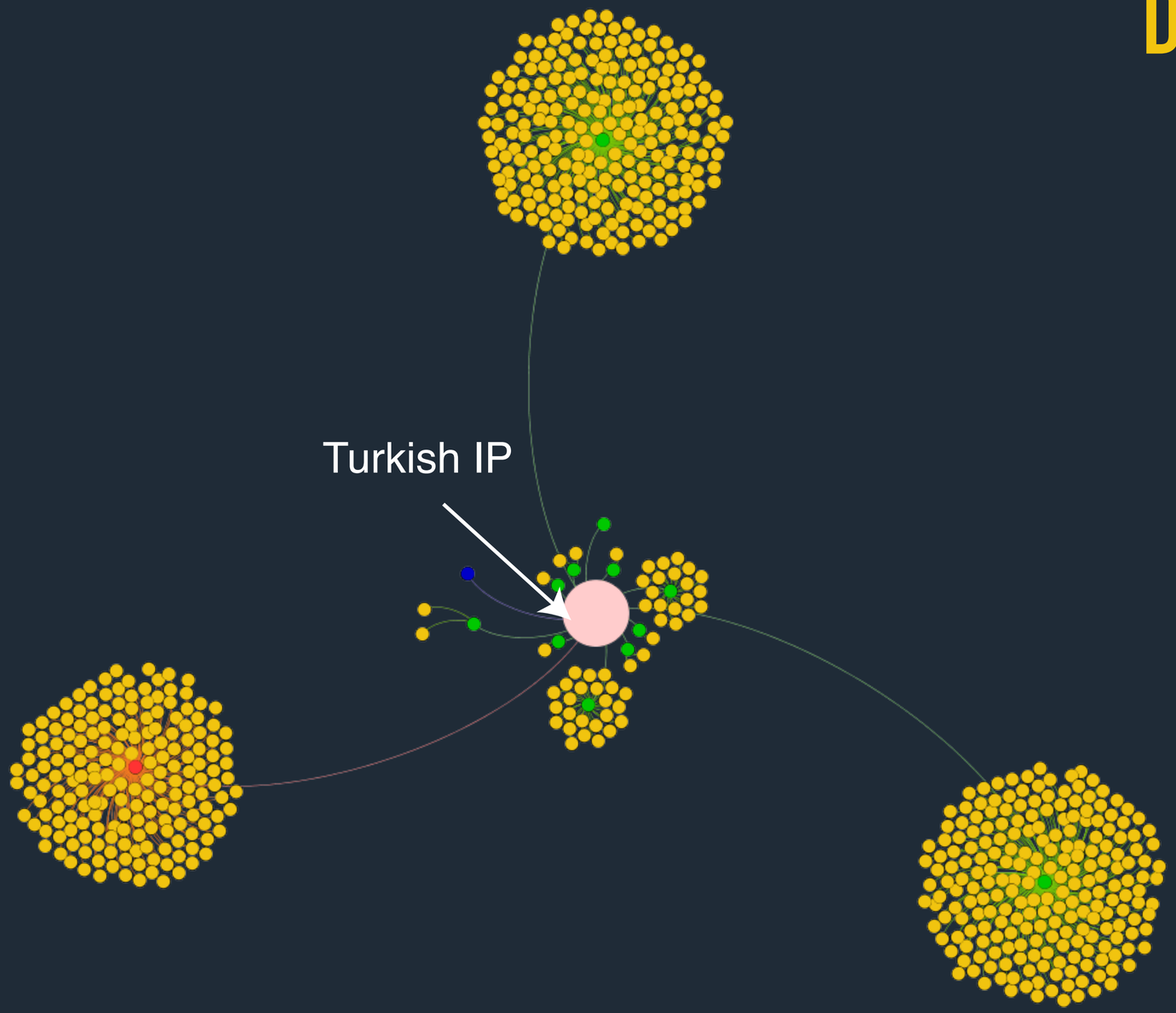
Torrent Correlation

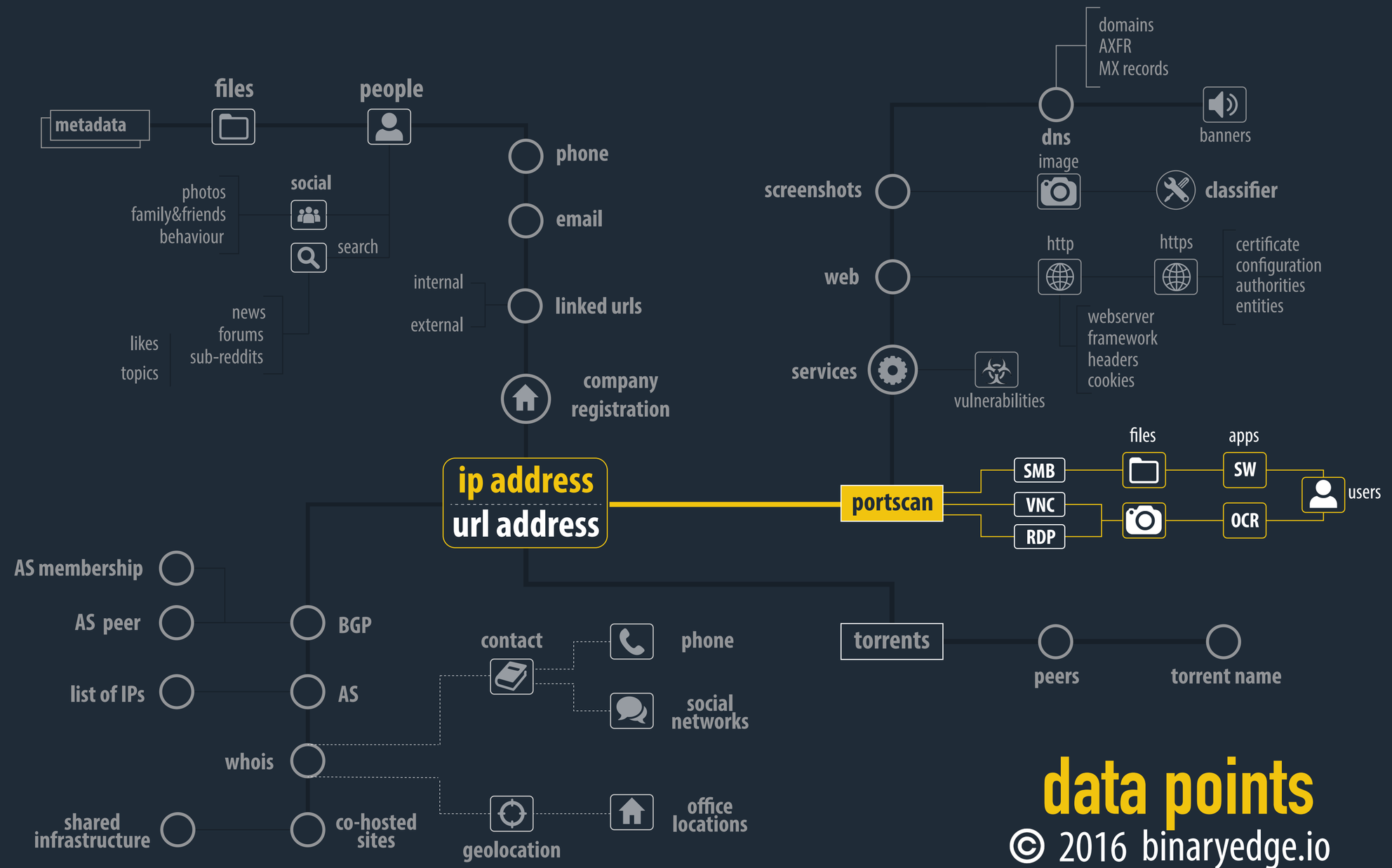


Data correlation



Data correlation

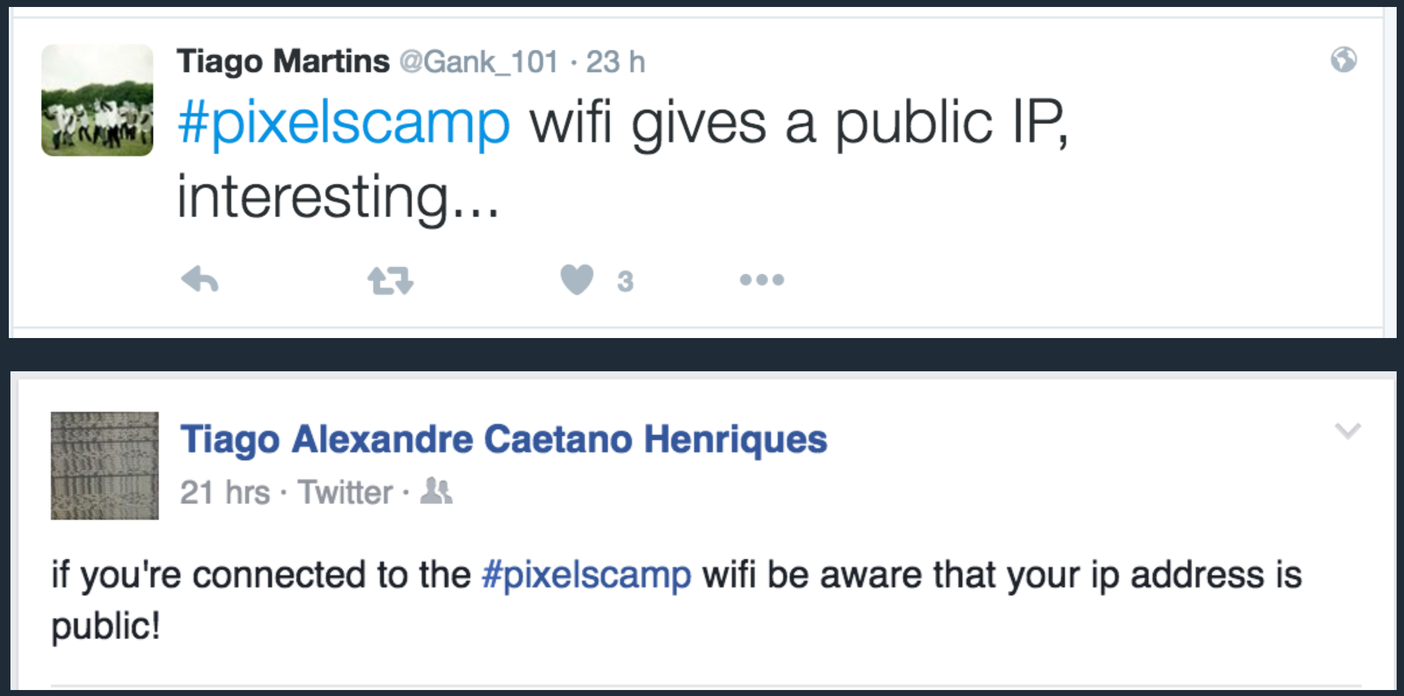



files
people
metadata
phone
email
linked urls
company registration
internal
external
social
photos
family&friends
behaviour
search
news
forums
sub-reddits
likes
topics
domains
AXFR
MX records
dns
banners
image
screenshots
classifier
web
http
https
certificate
configuration
authorities
entities
webserver
framework
headers
cookies
services
vulnerabilities
files
apps
SMB
VNC
RDP
SW
OCR
users
portscan
ip address
url address
AS membership
AS peer
list of IPs
BGP
AS
whois
shared infrastructure
co-hosted sites
contact
phone
social networks
geolocation
office locations
torrents
peers
torrent name
data points
© 2016 binaryedge.io

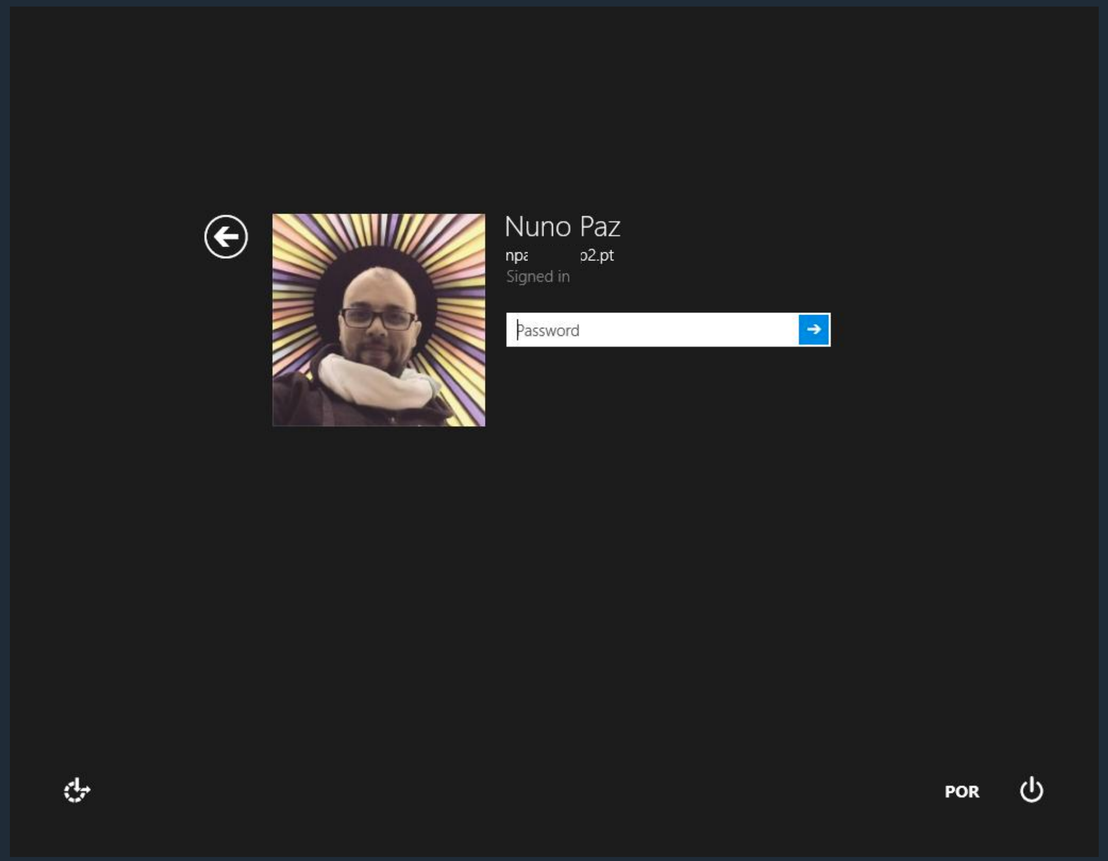

00:02:09

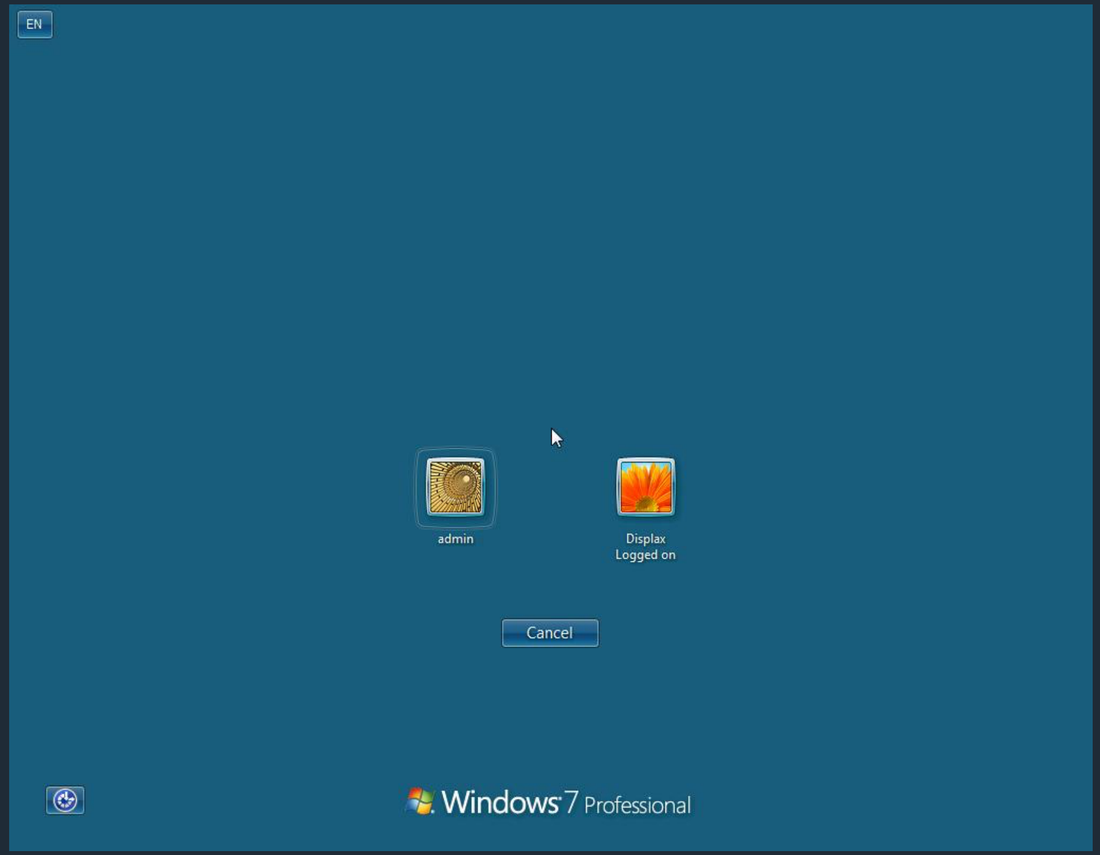
DEMO

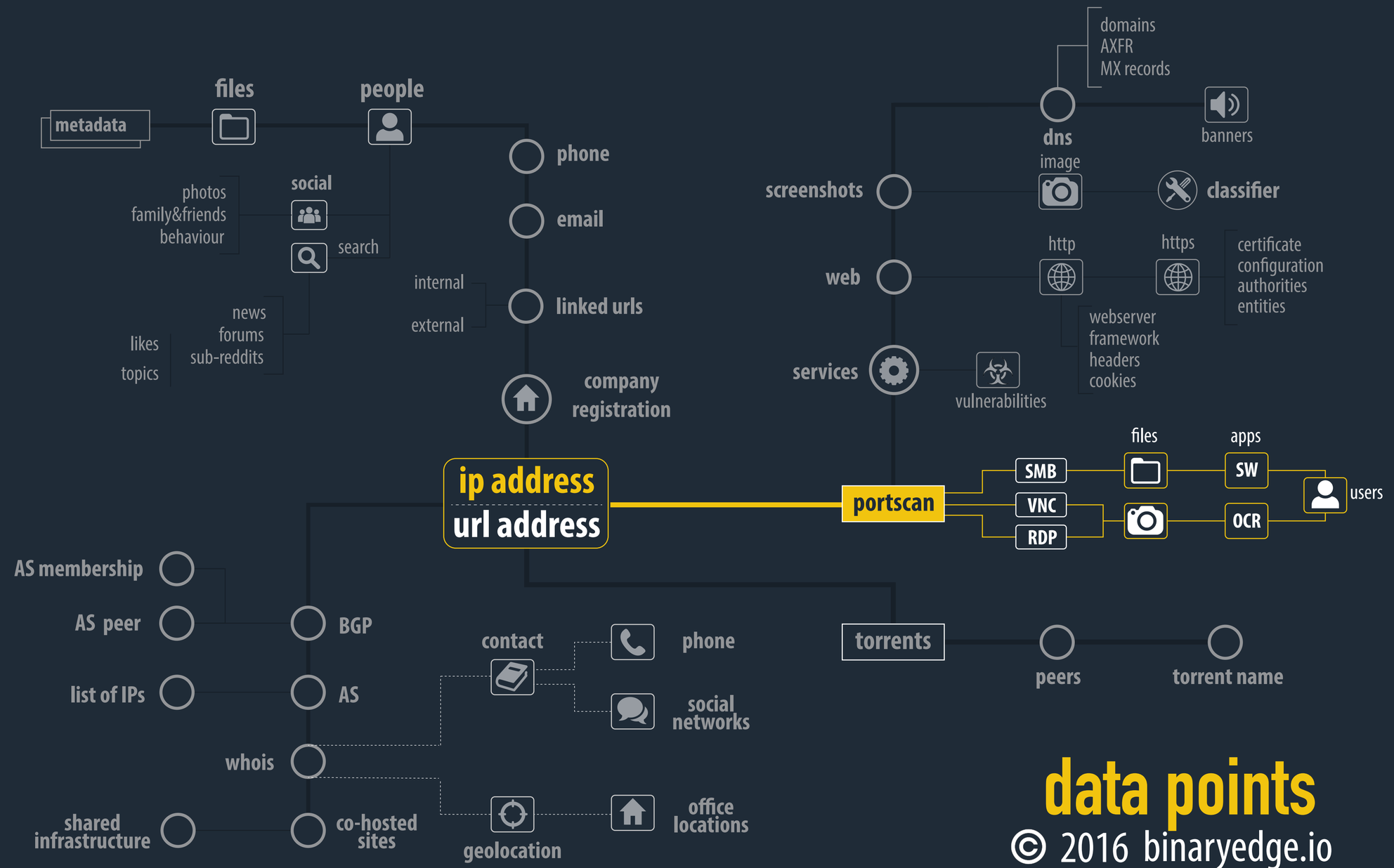
At PixelsCamp



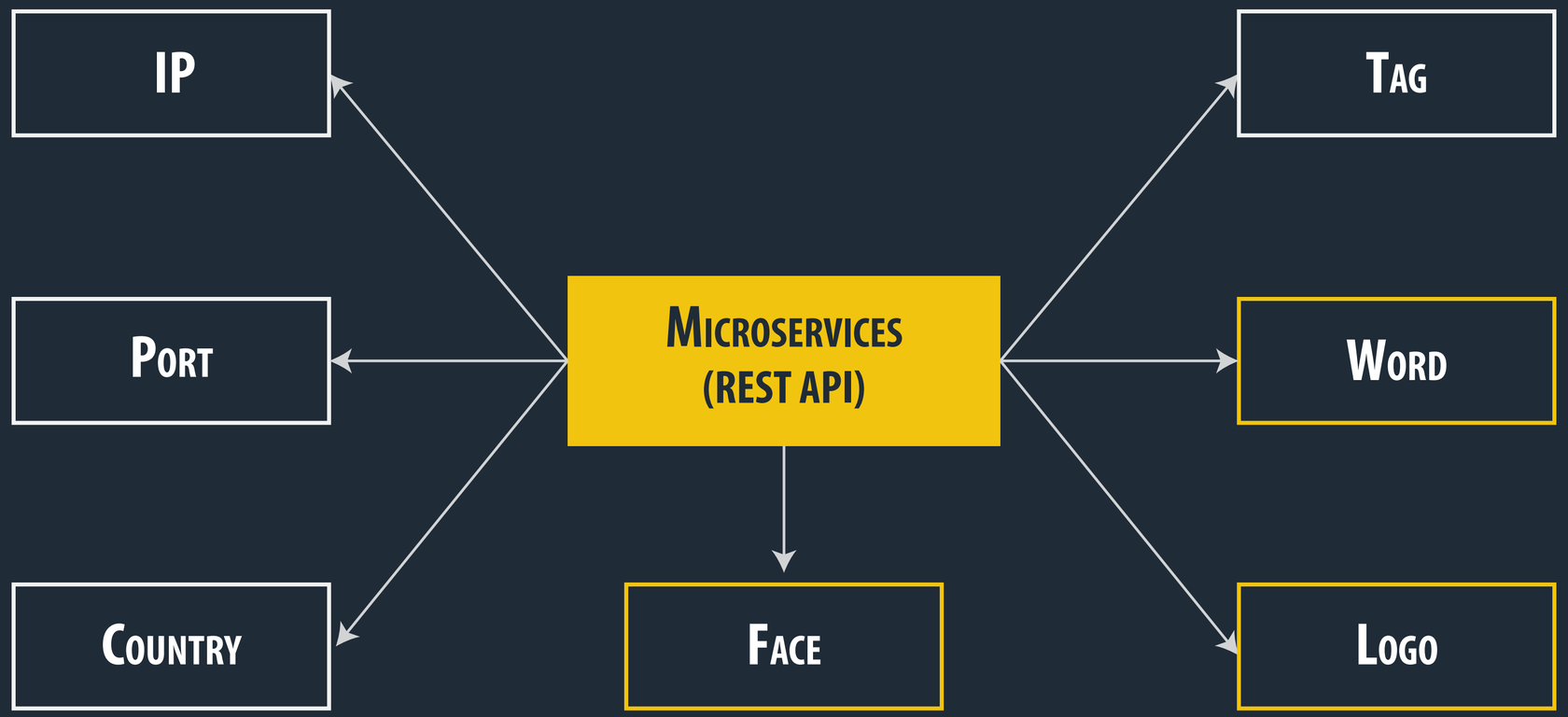
At PixelsCamp



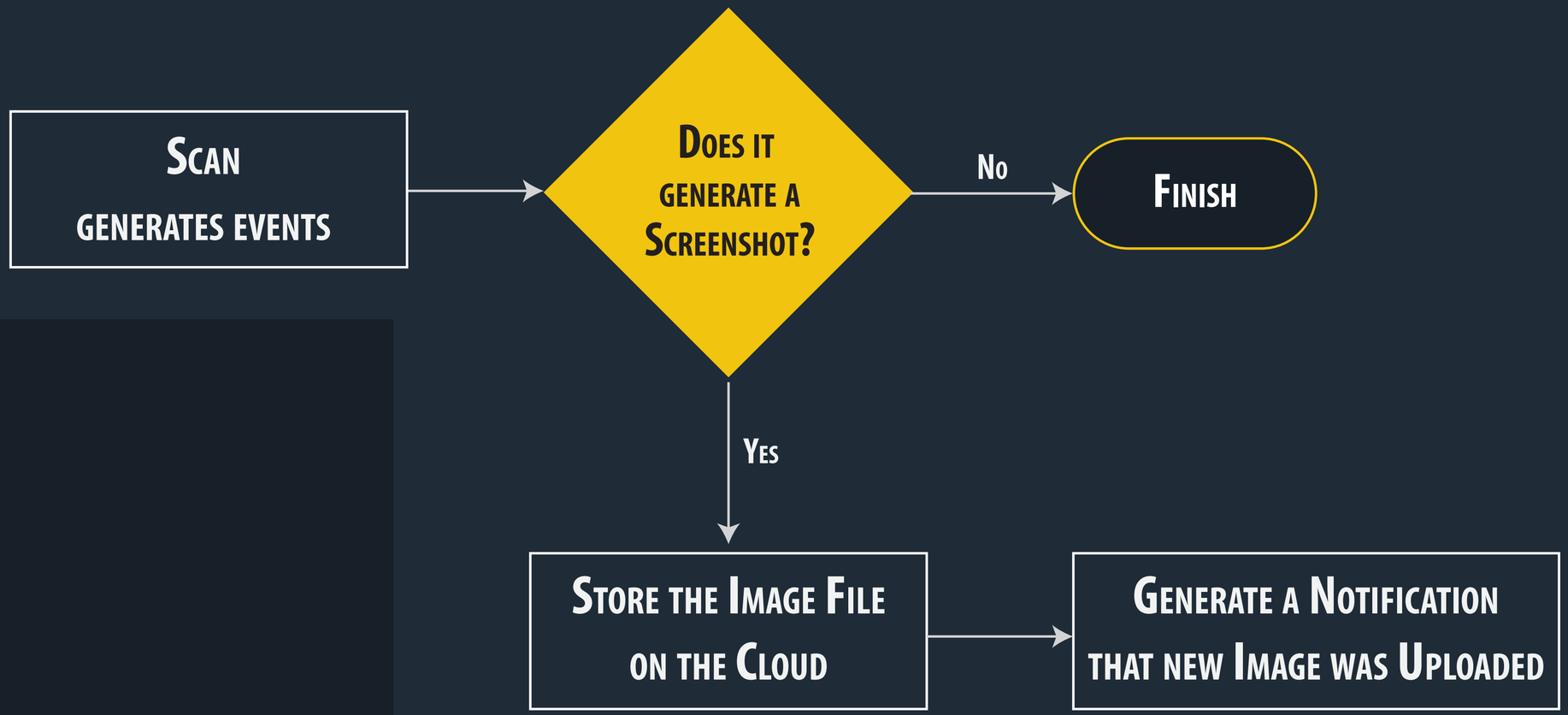



files
people
metadata
phone
email
linked urls
company registration
internal
external
social
photos
family&friends
behaviour
search
news
forums
sub-reddits
likes
topics
domains
AXFR
MX records
dns
banners
image
screenshots
classifier
web
http
https
certificate
configuration
authorities
entities
webserver
framework
headers
cookies
services
vulnerabilities
files
apps
SMB
VNC
RDP
SW
OCR
users
portscan
ip address
url address
AS membership
AS peer
list of IPs
BGP
AS
whois
shared infrastructure
co-hosted sites
contact
phone
social networks
geolocation
office locations
torrents
peers
torrent name
data points
© 2016 binaryedge.io

Microservices (REST API)



Scan




```
{
 "origin": {
  "type": "vnc",
...
 },
 "target": {
  "ip": "XX.XXX.XX.XXX",
  "port": 5900
 },
 "result": {
  "data": {
   "version": "3.7",
   "width": "1366",
   "height": "768",
   "auth_enabled": false,
   "link": "https://5723981752938cbafeefbcfab42342342.jpg"
  }
 },
 "@timestamp": "2016-04-22T14:53:02.377Z"
}
```

Image Workflow

Image Workflow

Get image

Extract target metadata

Create image signature

Store data

Does it contain any content?

Yes

Enhance image for logo and face detection and OCR extraction

Perform logo and face detection and OCR extraction

Store results

Perform additional actions

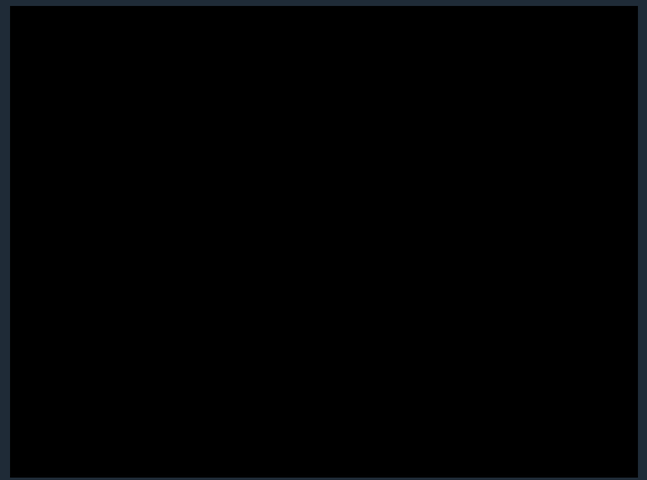
No

Finish

Filter

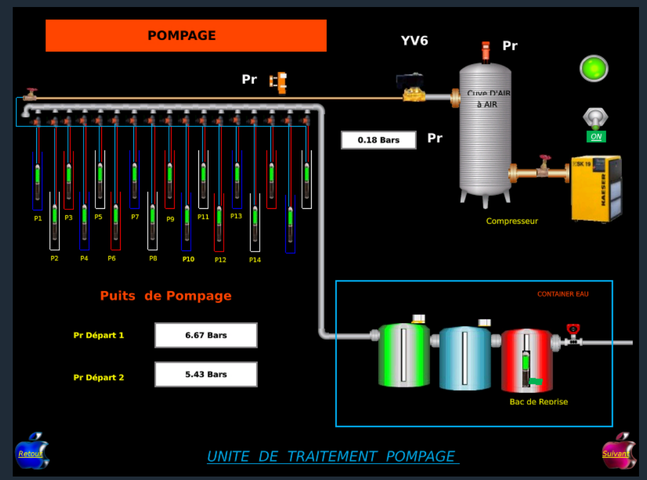
Shannon's Entropy

$$H(A) = -\sum_{i=1}^{n} p_i \log_2 p_i$$



Entropy = 0.00 bits

Entropy ~ 0.03 bits



Entropy ~ 2.13 bits

DEMO

Data Visualization

"a multidisciplinary recipe of art, science, math, technology, and many other interesting ingredients."

Andy Kirk, "Data Visualization: a successful design process"

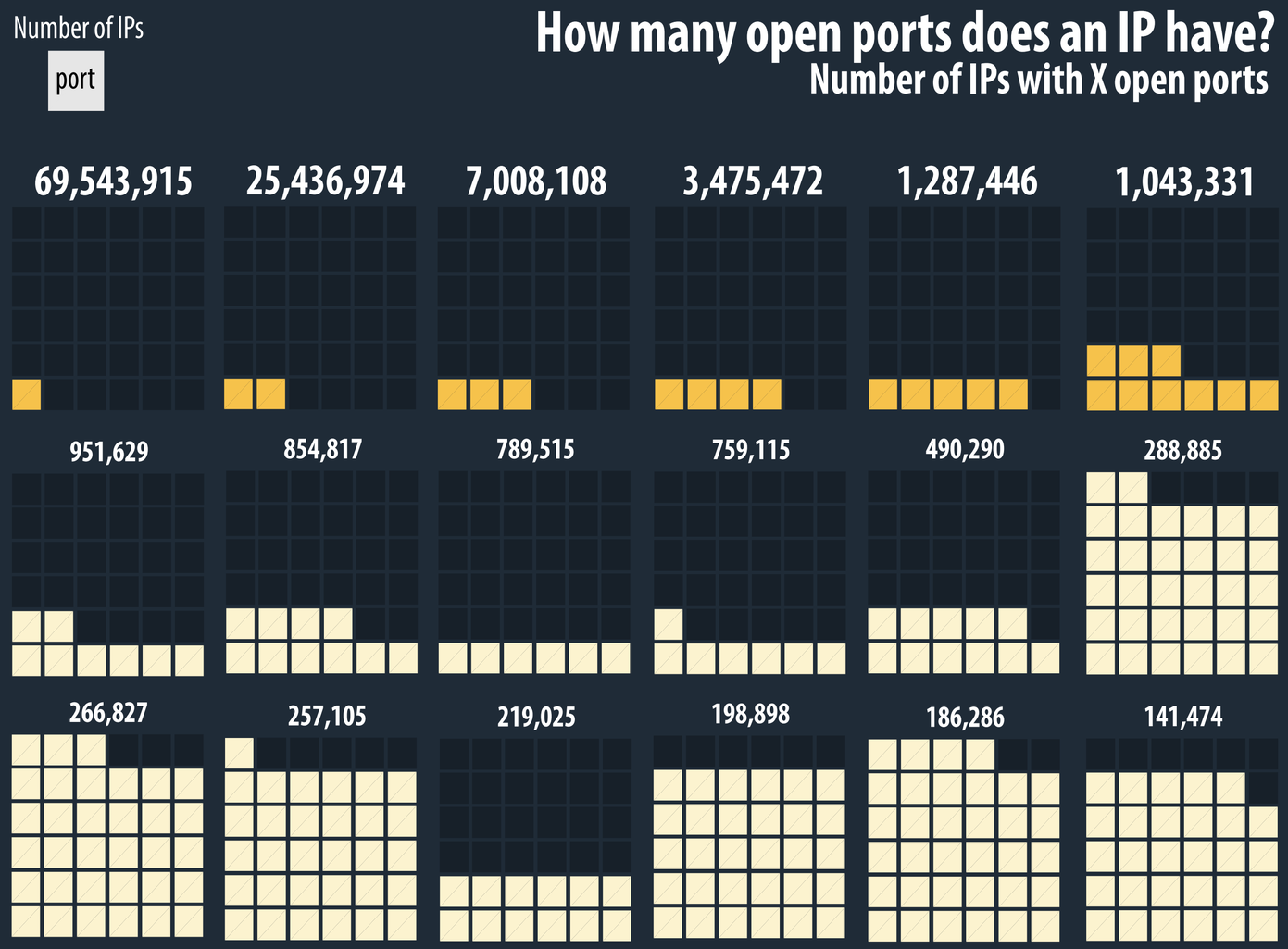
Data Visualization

Exploration → **Representation** → Details → Tools → Finishing Up

Experimentation is important

↓

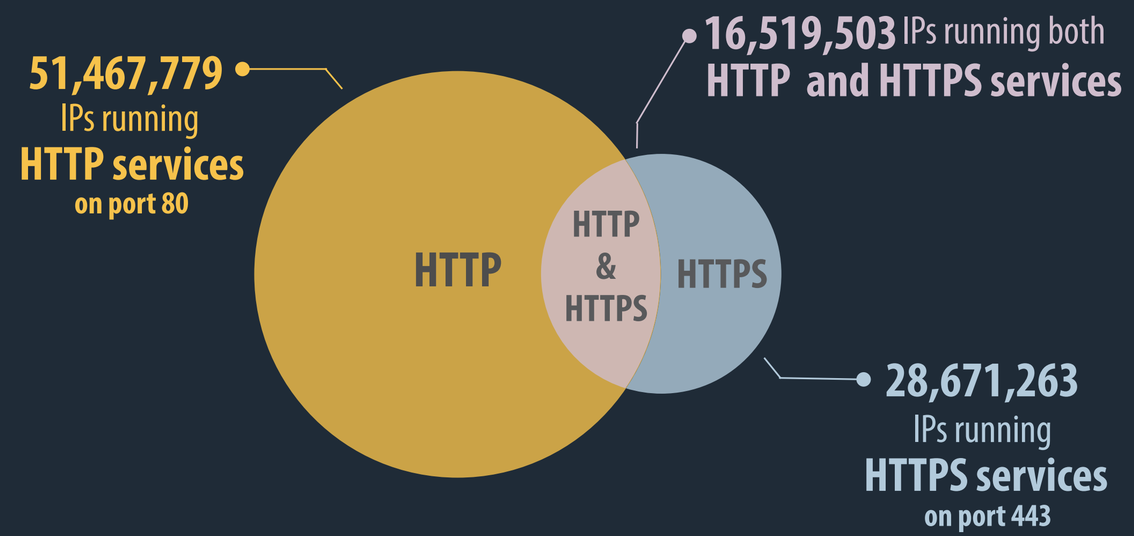
design can be used in the future



Data Visualization

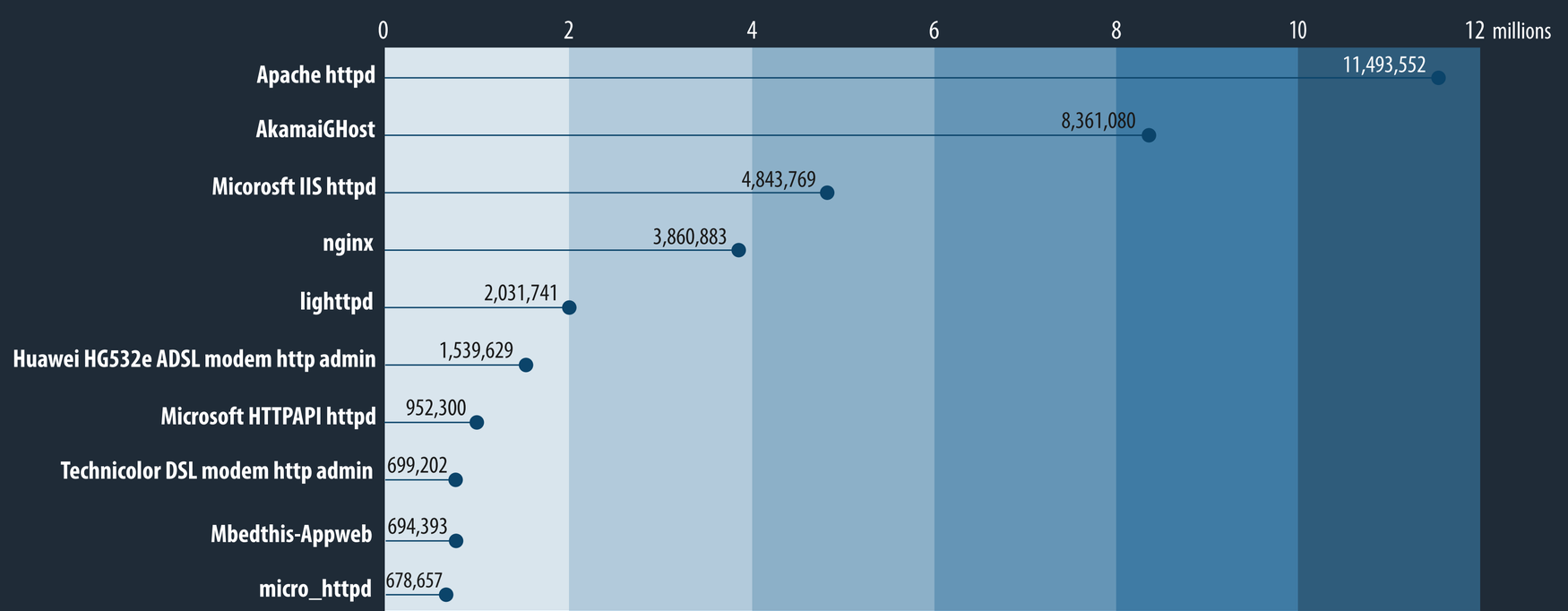

```
{
 "origin": {
  "type": "service-simple",
...
 },
 "target": {
  "ip": "XX.XX.XXX.XXX",
  "port": 80,
  "protocol": "tcp"
 },
 "result": {
 ...
    "service": {
     "product": "Microsoft HTTPAPI httpd",
     "name": "http",
     "extrainfo": "SSDP/UPnP",
     "cpe": [
      "cpe:/o:microsoft:windows"
     ]
   }
 },
 "@timestamp": "2016-04-22T04:07:18.161Z"
}
```


Distribution of IP addresses running encrypted and unencrypted services



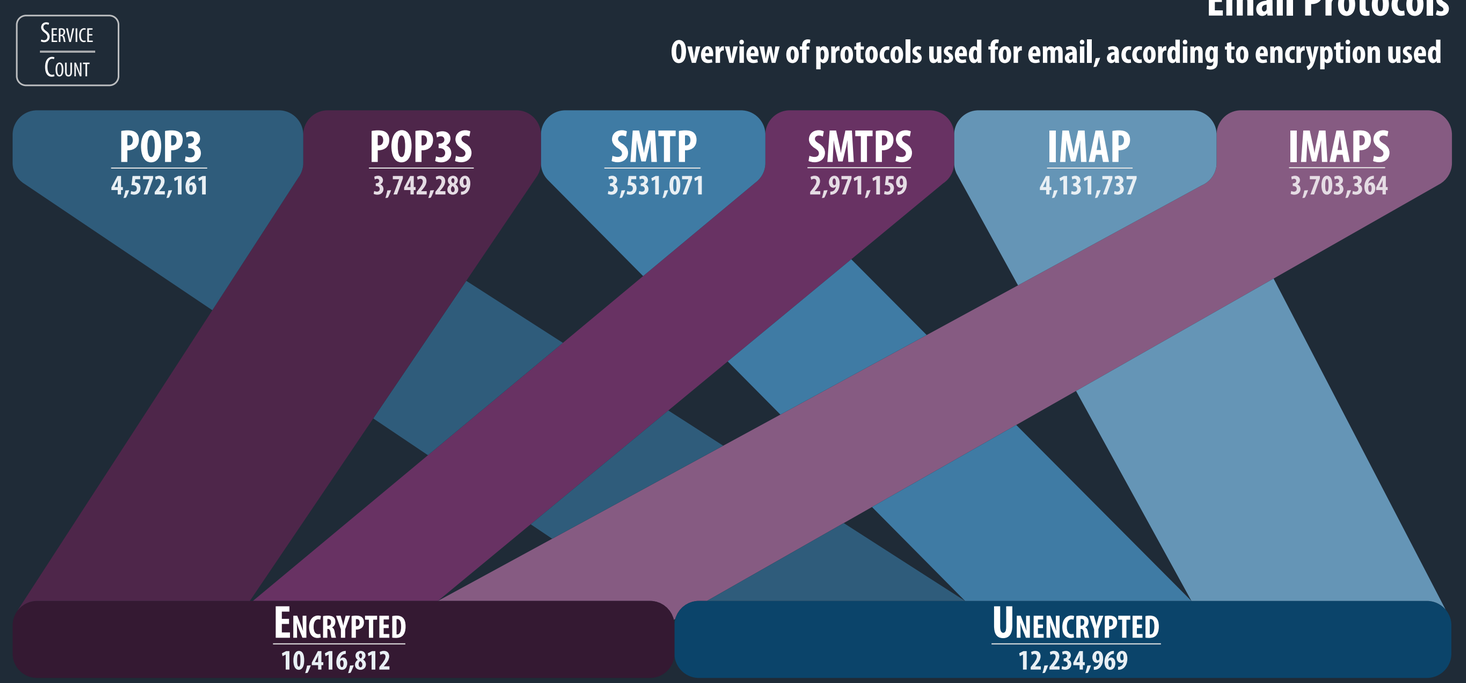
Data Visualization


```
{
...
  "result": {
    "data": {
      "apps": [
        {
          "name": "Apache",
          "confidence": 100,
          "version": "2.2.26",
          "categories": [
            "web-servers" ]
...
        }
      }
    }
  }
```



Data Visualization


```
{
 "origin": {
  "type": "service-simple",
...
 },
 "target": {
  "ip": "XX.XXX.XXX.XX",
  "port": 143,
  "protocol": "tcp"
 },
 "result": {
...
    "service": {
     "method": "probe_matching",
     "product": "Dovecot imapd",
     "name": "imap",
     "cpe": [
      "cpe:/a:dovecot:dovecot"
     ]
...
 },
 "@timestamp": "2016-04-22T01:56:54.583Z"
}
```

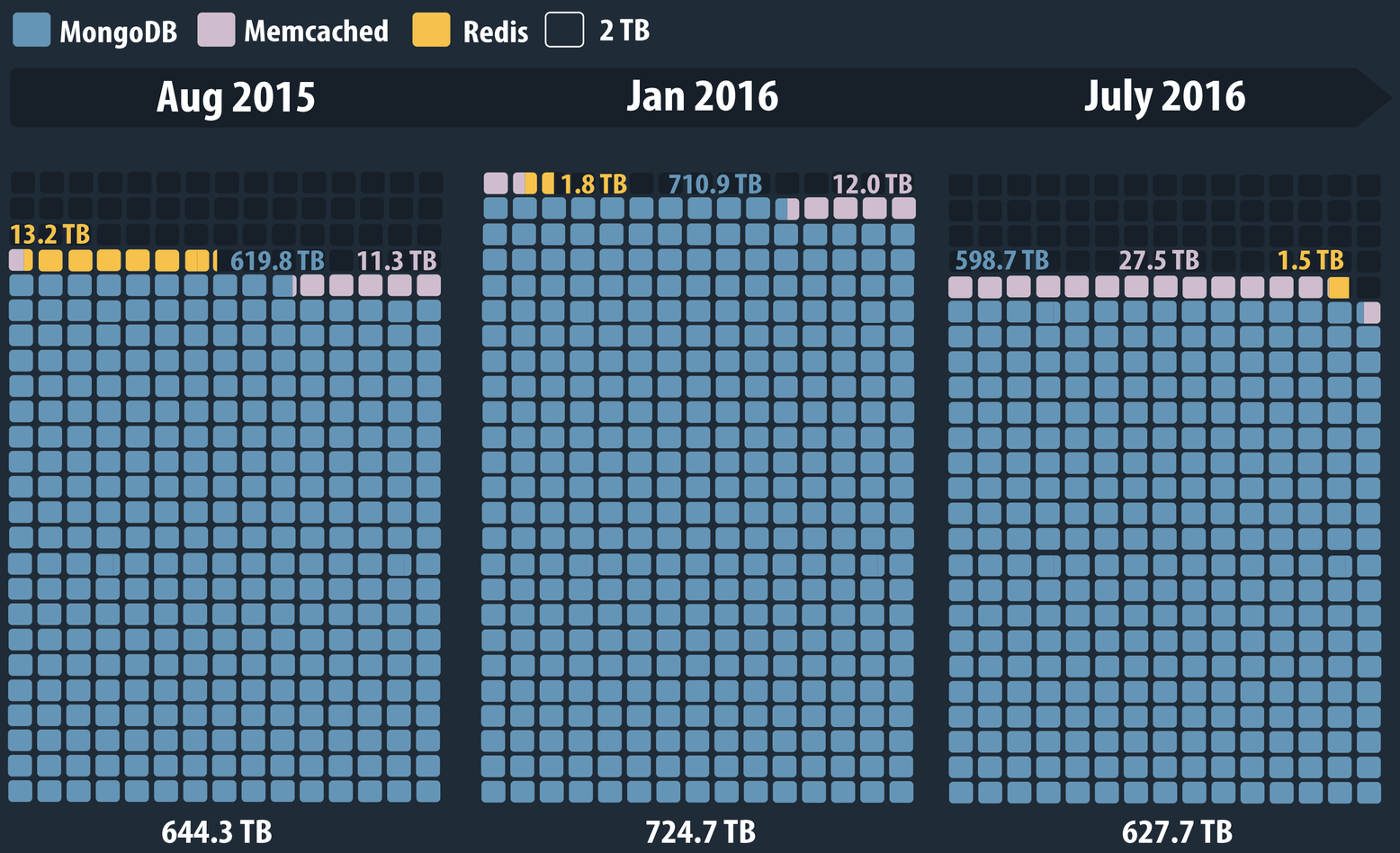



Data Visualization


```
{
  "origin": {
    "type": "redis",
...
  },
  "target": {
    "ip": "XXX.XX.XX.XXX",
    "port": 6379
  },
  "result": {
    "data": {
      "redis_version": "3.0.6",
...
      "used_memory": 1374760,
      "used_memory_human": "1.31M",
      "used_memory_rss": 1839104,
      "used_memory_peak": 25195656,
      "used_memory_peak_human": "24.03M",
      "used_memory_lua": 36864,
      "mem_fragmentation_ratio": 1.34,
...
  },
  "@timestamp": "2016-04-22T15:37:10.913Z"
}
```


Big Data Technologies

Changes in amount of data exposed without security

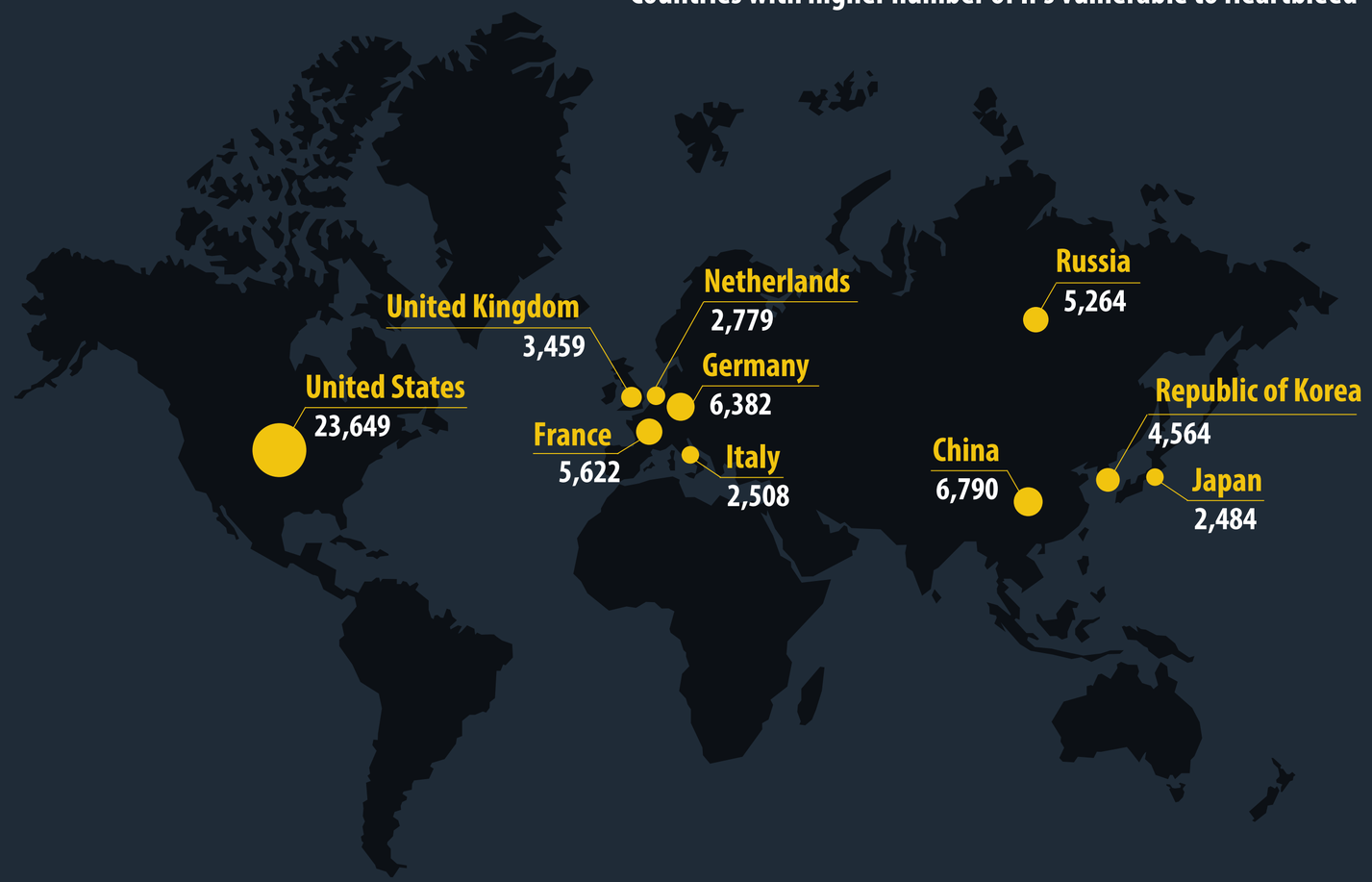


Data Visualization


```
{
 "origin": {
  "type": "ssl",
 },
 "target": {
  "ip": "XXX.XX.X.XXX",
  "port": 443
 },
 "result": {
  "data": {
   "vulnerabilities": {
    "heartbleed": {
     "is_vulnerable_to_heartbleed": true
    },
    "openssl_ccs": {
     "is_vulnerable_to_ccs_injection": false
    }
   },
  }
 }
}
```

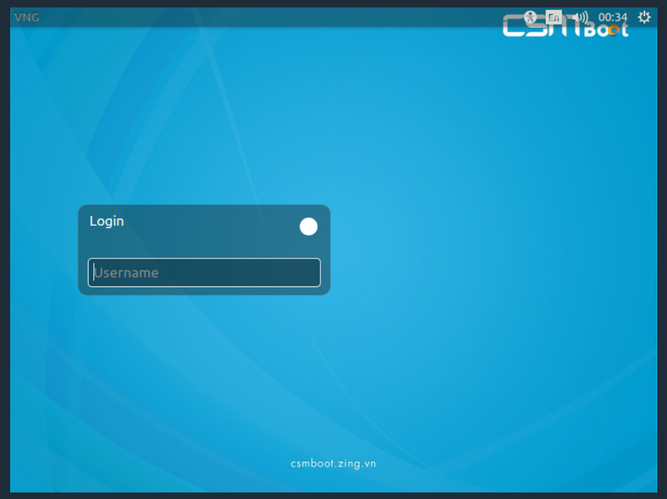
Heartbleed

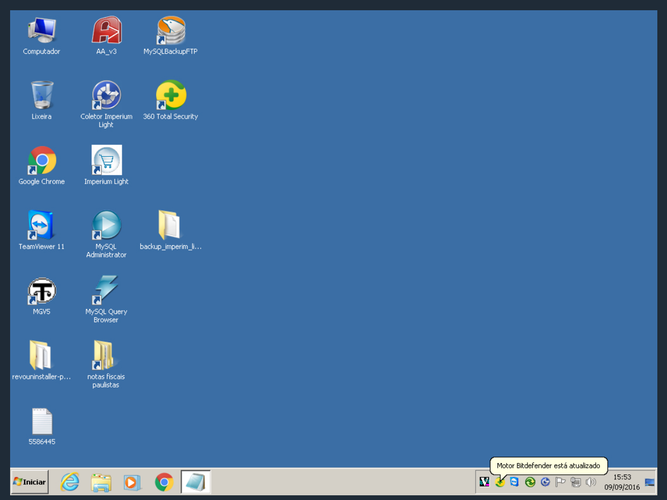
Countries with higher number of IPs vulnerable to Heartbleed



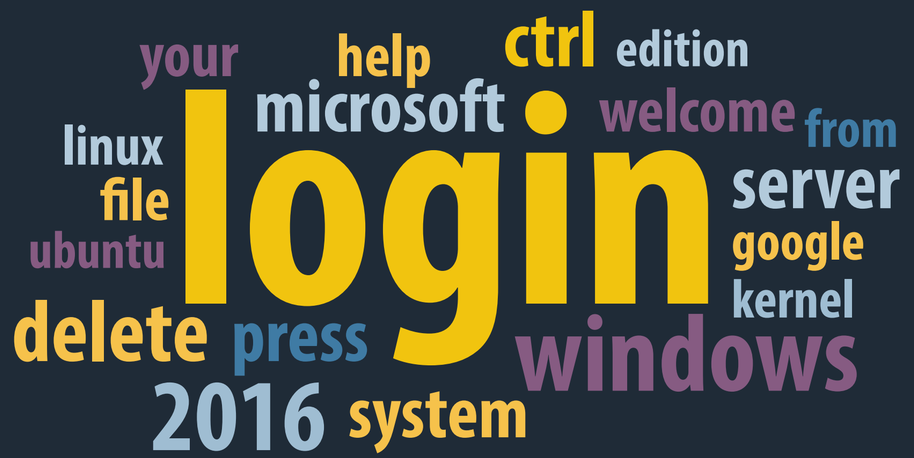
Data Visualization

EXPLORATION → **REPRESENTATION** → DETAILS → TOOLS → FINISHING UP



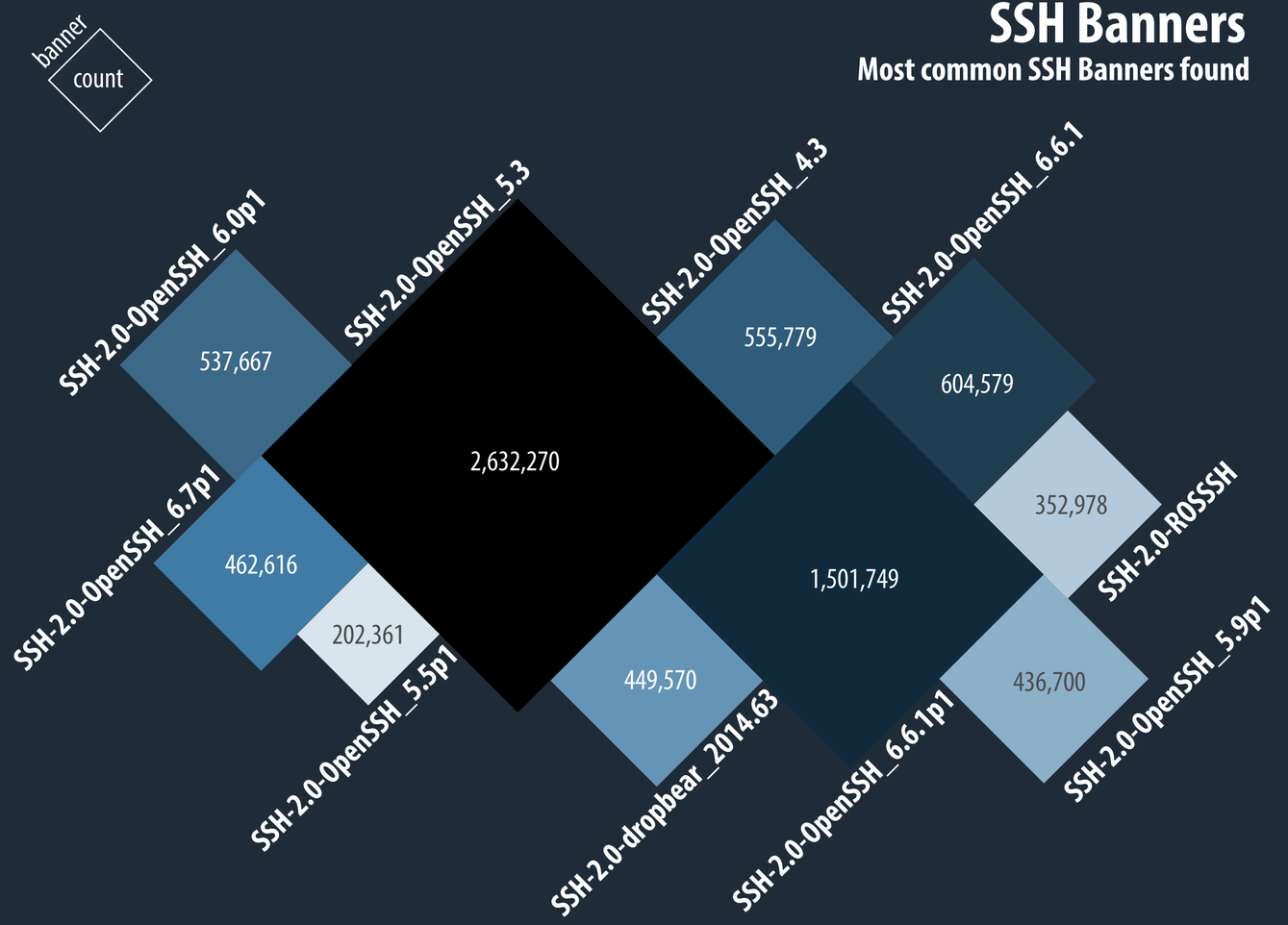


VNC wordcloud



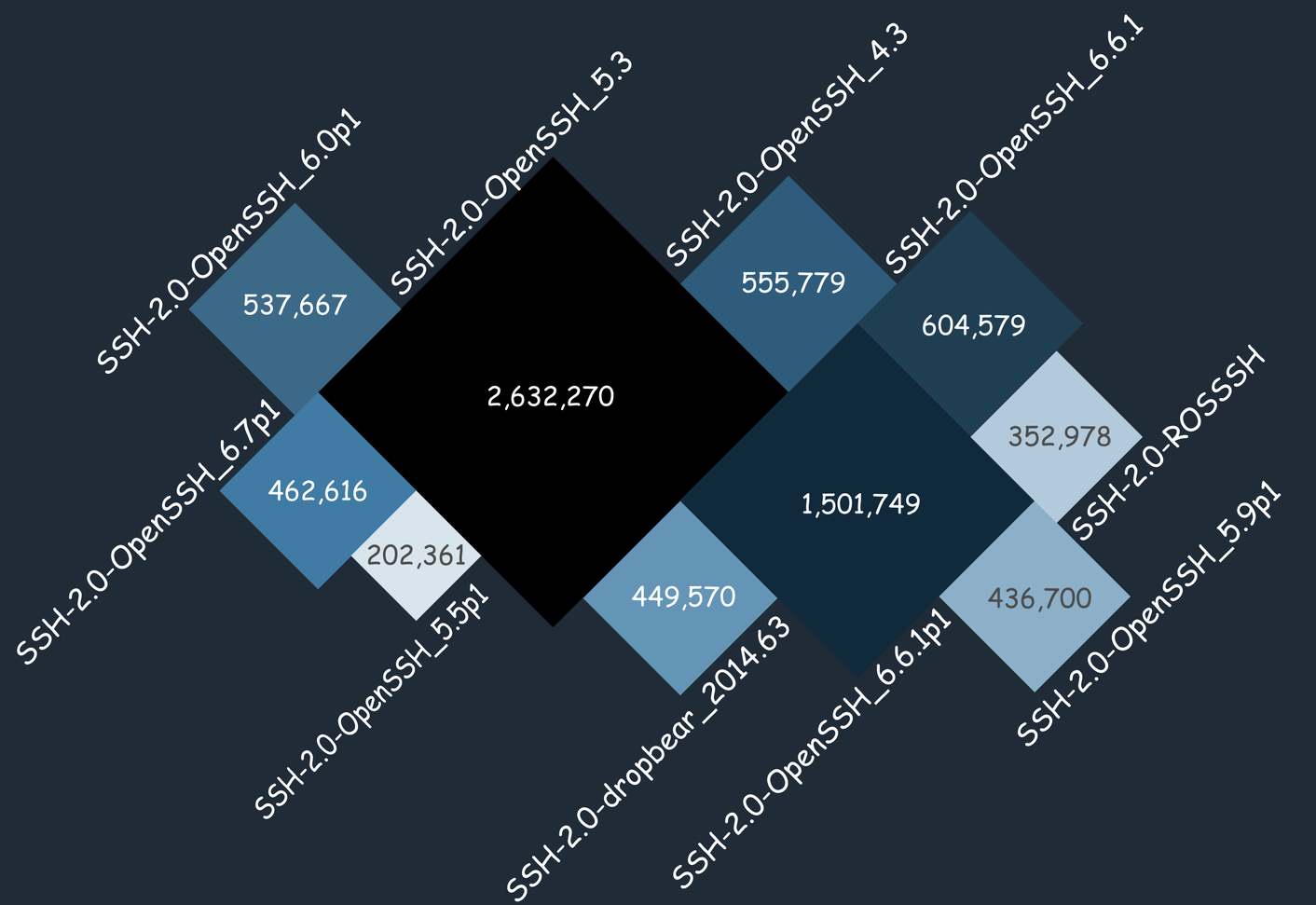
Data Visualization


```
{
 "origin": {
  "type": "ssh",
  "job_id": "client-816f1185-4bc1-4b5f-9a7d-61a2df315a6b",
  "client_id": "client",
  "country": "uk",
  "module": "grabber",
  "ts": 1453385574412
 },
 "target": {
  "ip": "X.X.X.X",
  "port": 22,
  "protocol": "tcp"
 },
 "result": {
  "data": {
...
   "banner": "SSH-2.0-OpenSSH_6.6.1p1"
  }
 }
}
```

Data Visualization


```
{
 "origin": {
  "type": "ssh",
  "job_id": "client-816f1185-4bc1-4b5f-9a7d-61a2df315a6b",
  "client_id": "client",
  "country": "uk",
  "module": "grabber",
  "ts": 1453385574412
 },
 "target": {
  "ip": "X.X.X.X",
  "port": 22,
  "protocol": "tcp"
 },
 "result": {
  "data": {
...
   "banner": "SSH-2.0-OpenSSH_6.6.1p1"
  }
 }
}
```

Data Visualization


```
{
  "origin": {
...
  },
  "target": {
    "ip": "X.X.X.X",
    "port": 22,
    "protocol": "tcp"
  },
  "result": {
...
      {
        "cypher": "ssh-rsa",
        "key": "AAAAB3NzaC1yc2EAAAABIwAAAQEAudfUFJtWp8R5qPxXB0acGHctH0Yyx-
VrZZfvnG37osNc32kX35aXVm8Ulk49zl/jMIIQnzP7zeOUJeJJsyXsG6Cu3qjLvD5qlc0tRjoV
mV08aDgAsfeq7qQFEzzDqyoL8kV9akj8WyP+aN3QHvM4a/+3Y+UTVqrw5jSUiIIW5JOd+
UWzSz6SCGalFbop1wGELUTY6MDTHwwn+qXYgltQG6hP5tI9tl3gAVajIHg2IxM8IXz4SYH
33ZeOPypzrcr1/DvFx1s0773eGSArIi83BeYyxvN/T68RxIqAieLxVy8zJgyevpqHpUX7/+kDu
vVZdfKkmFoNzBTEiIvR5eMrjTw==",
        "fingerprint": "5b:71:c9:85:6a:ea:40:dc:62:95:4c:25:40:b7:97:55",
        "length": 2048
      }
    ],
...
    }
  }
}
```


SSH Key Lengths

Most common key lengths found

Key length	count
1024	5,068,711
2048	3,740,593
1040	641,719
1032	186,070
4096	13,845
512	9,064
2056	7,830
2064	6,265
1016	6,212
768	4,755

Data Visualization

Tools

Automation

- Programming Language to create plots
- Fine tunning in illustrator (make it better for the audience)
- Hand-editing process
- Human error

BALANCE

Originality

- Automated Analysis
- Illustrator (or other tool) to create visualization solution
- Human error

Data Visualization

Document every step of the process

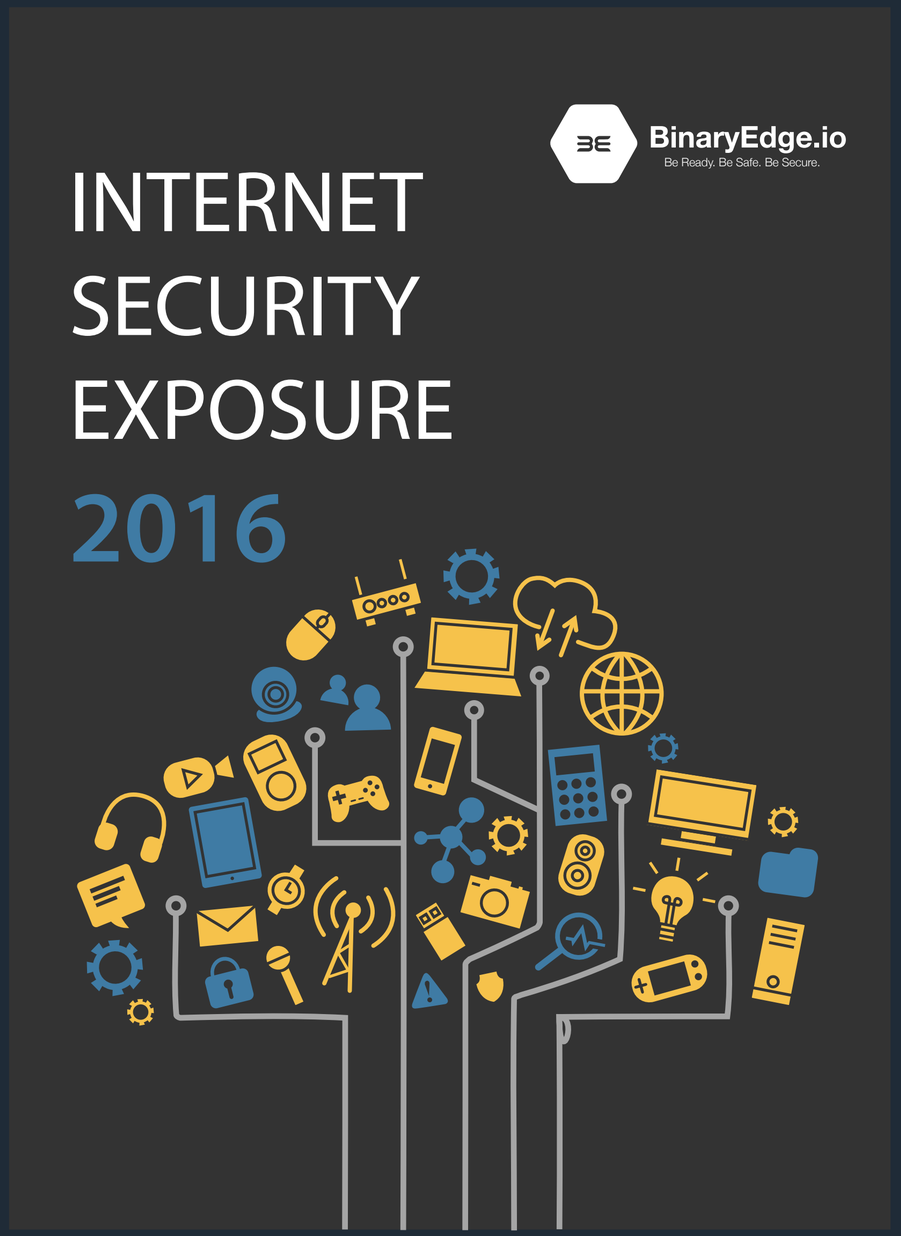
- Calculations
- Choices of visualisations
- Choices of data points

Review Everything

- What could have been done differently?
- What could be better?

Take Constructive Feedback

- Even if it means to start over
- A visualization can be used in the future



ise.binaryedge.io

#1
2016

THE SCIENCE BEHIND THE DATA

CREATED BY
BINARYEDGE