# Find A Hay In Haystack! How Splunk Help Recruit To Detect 0.000001% Threat And More...


Mitsuhiro Nakamura

Senior Security Engineer, Recruit Technologies Co., Ltd


# Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

# Agenda

Bio & About Company

1. Heuristic Analysis Using Splunk
2. Data Visualization Using Splunk D3.Js Apps
3. Deep Learning (Extra Challenge)

# Bio

## Mitsuhiro Nakamura (hiro)

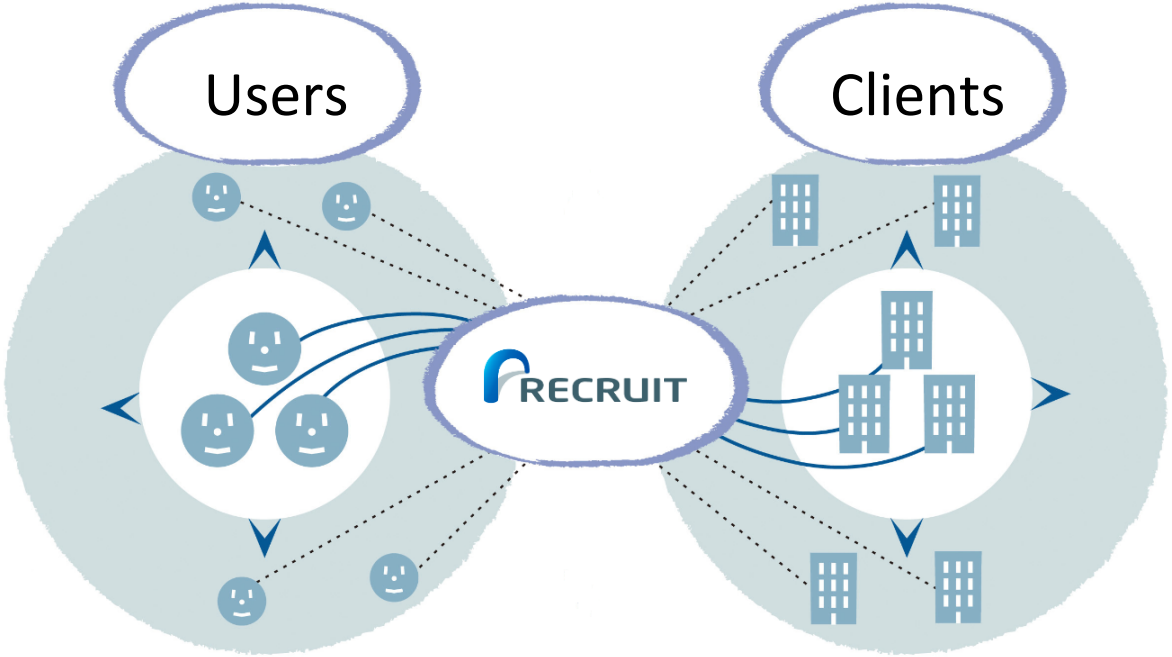
- 10+ years experience in cyber security specializing Pen-Test, Forensics, and Incident-Response.
- Built Web Application Vulnerability Assessment Methodology.
- Data analysis for Threat Detection using Splunk.

## like: Windsurfing (Wave)

# Recruit Holdings Co., Ltd.

| | |
|---|---|
| Founded | **1960** |
| Employees | **38,000** |
| IPO | **Oct/2014** |
| Annual Sales | **$ 14.0 bn** |

# Business Model




Housing and Real Estate
JPY **83.9Bn**

Bridal
JPY **53.6Bn**

Travel
JPY **53.4Bn**

Dining
JPY **34.3Bn**

Beauty
JPY **39.9Bn**

Domestic Recruiting
JPY **239.8Bn**

Domestic Staffing
JPY **389.5Bn**

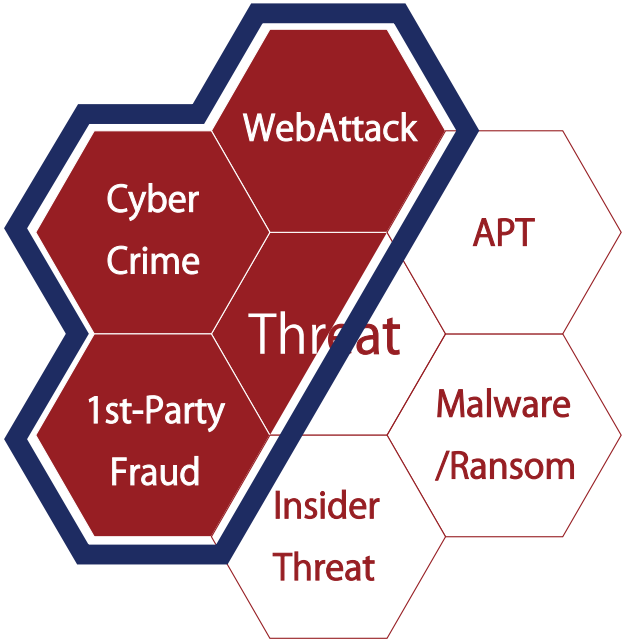
We are committed to creating chances to discover “Opportunities for Life,” connecting users and clients through new channels.

# Domestic


# Abroad (M&A)


200+ Web Commercial Sites

# What Is “Threat” On The Recruit?




**Production**

<domestic>

- Users : Tens of millions

**Internal-LAN**

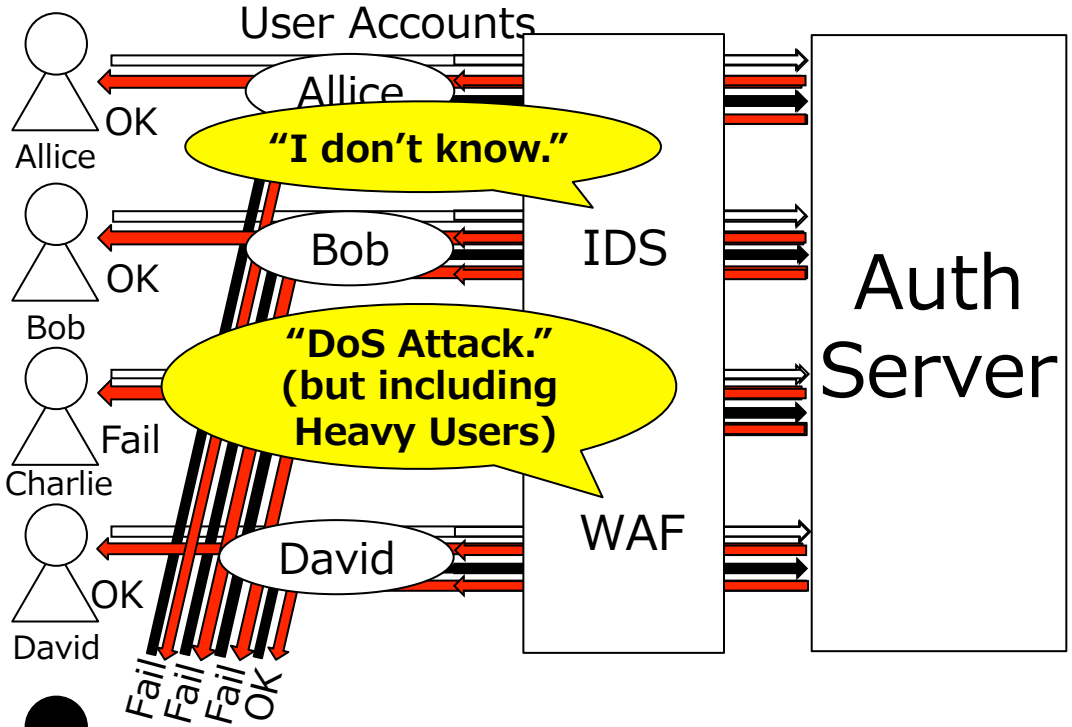
<domestic>

- Employees : 38 thousands

Today‘s my topic is..

**“List-Type Account Hacking”**

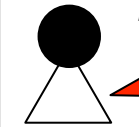
# What Is The “List-type Account Hacking”?




attacked at
Sep/2014

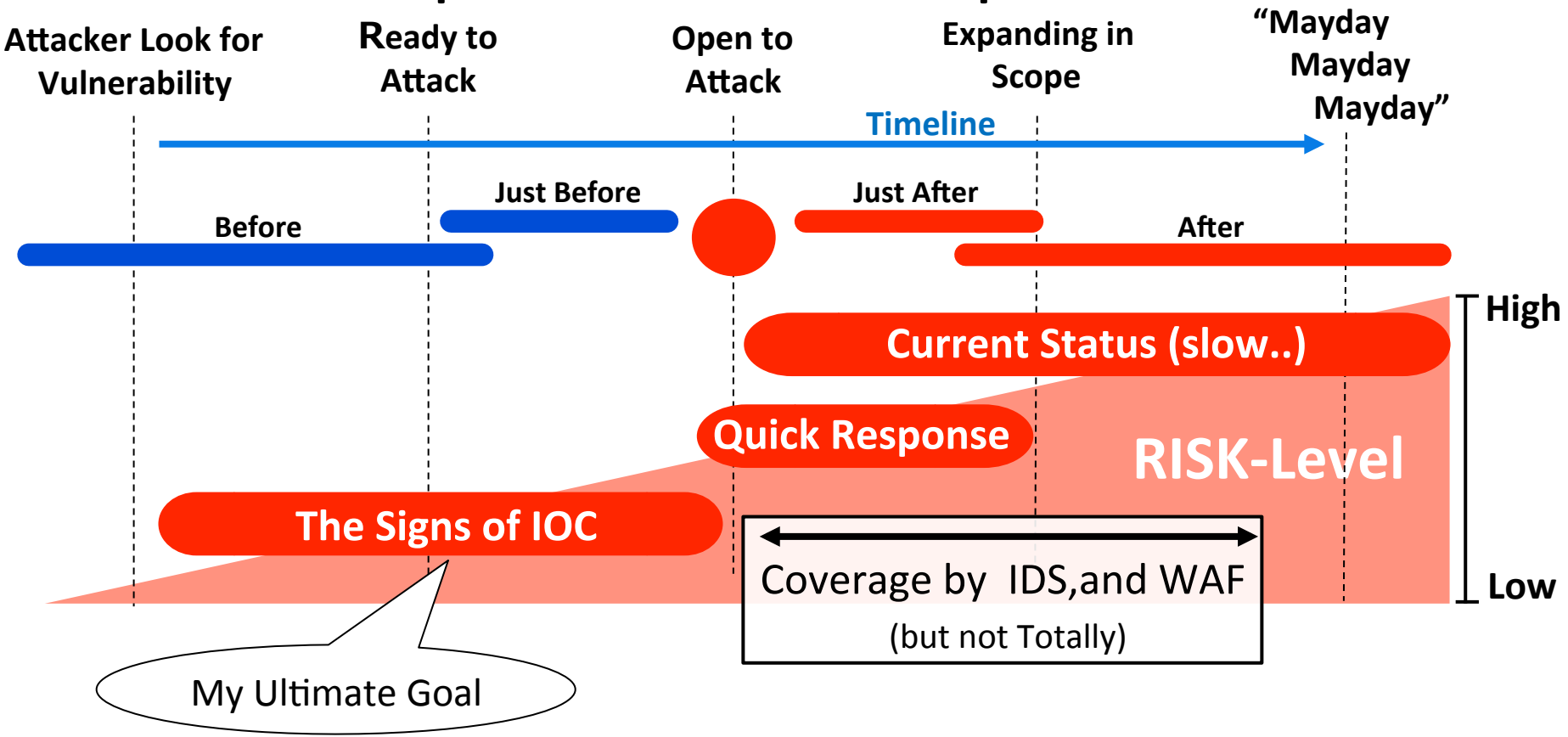
damaged
9749 ID

We were wild-samurai
at that time!


**Spoofed Login-Attempts using Their accounts.**
**Attackers has million Accounts-List.**
**thus Success Rate is higher than Simple-BruteForce-Attack.**

# Rapid Incident Response




# The Reason I Chose **Splunk**

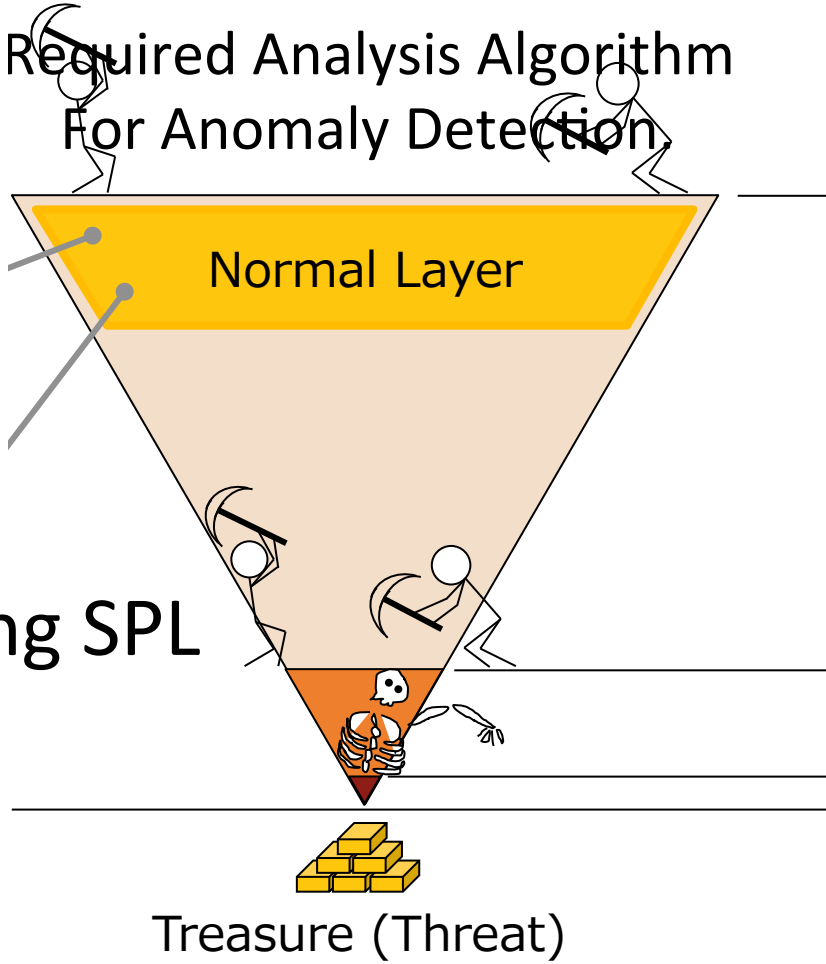
1. Log says Everything.

2. Real-time analysis Capability.

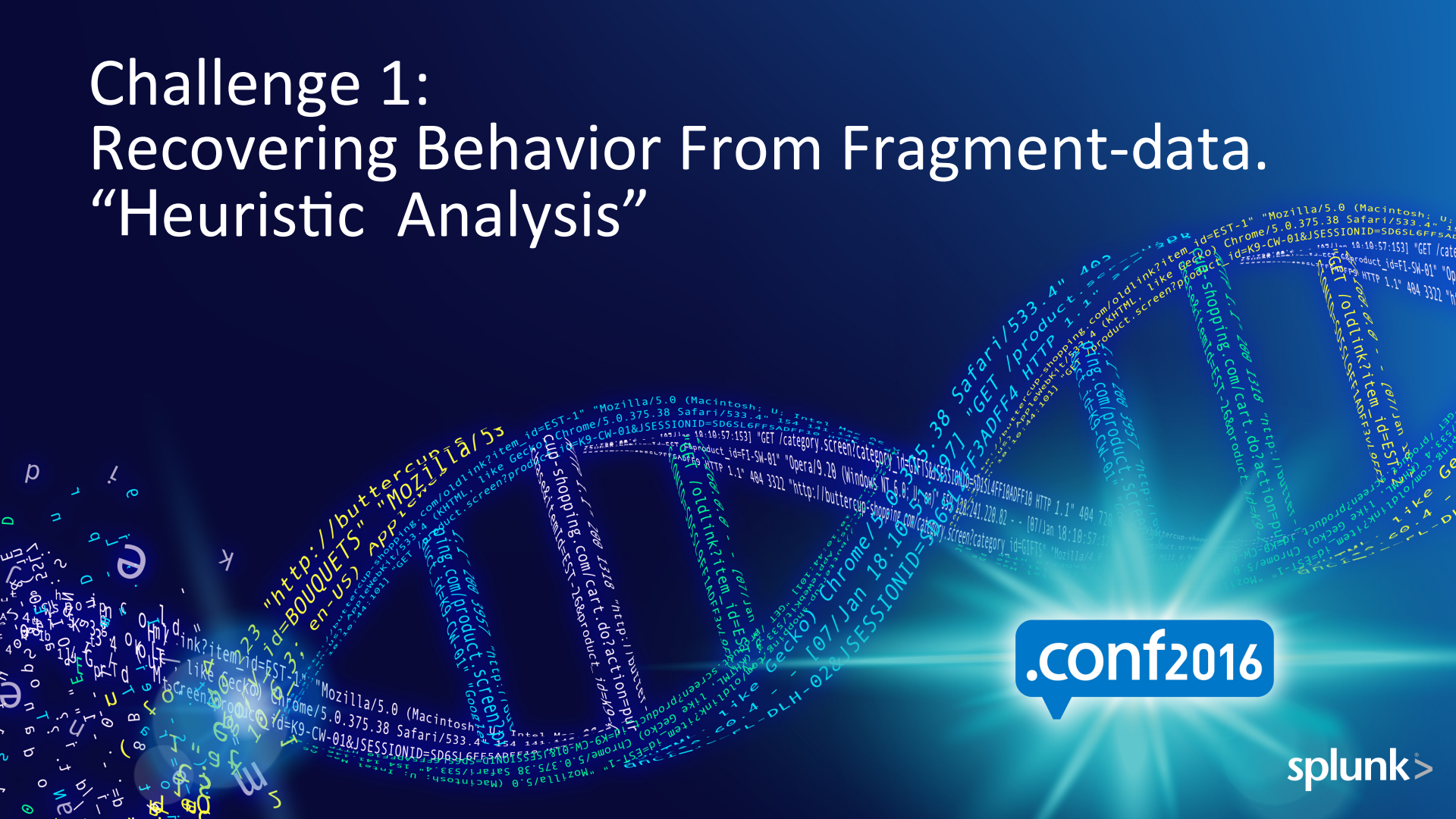
3. SPL.


# Required Analysis Algorithm For Anomaly Detection.

using SPL



# Challenge 1:
# Recovering Behavior From Fragment-data.
# “Heuristic  Analysis”

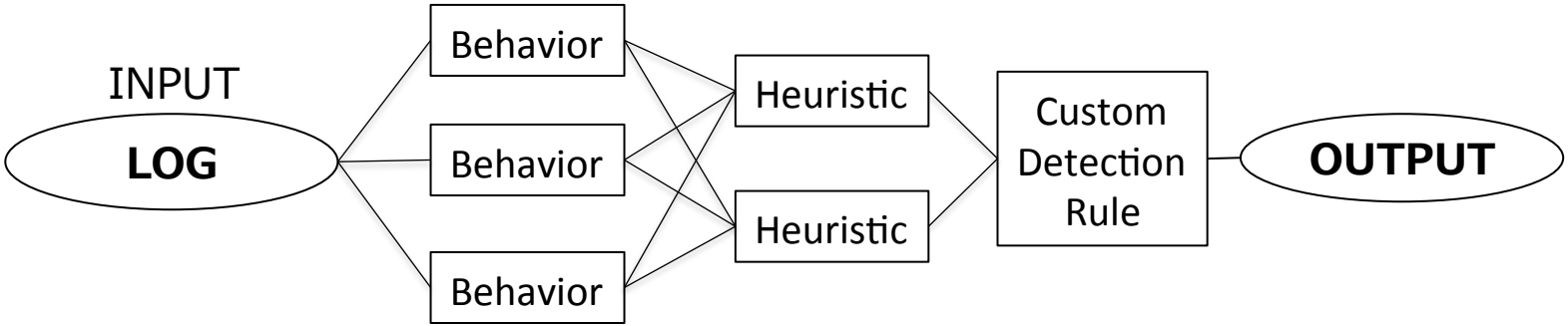


| Index | source | Source type | Key-Value Pair | By |
|---|---|---|---|---|
| Recruit | Apache | Apache | URL,HttpStatus, etc.. | SrcIP |
| | Tomcat | Tom_Act | All of Manipulations,etc.. | SrcIP,BrowserHash |
| | Auth-DB | DB_Login_Act | Results, RecruitID, UniqID | |
| Summary | ※ All of the Statistics from above | | | |

**& Lookup**

iplocation / dnslookup / Tor-exitnode (update per day with crontab)/InternalAPI

▼SPL (Heuristic )

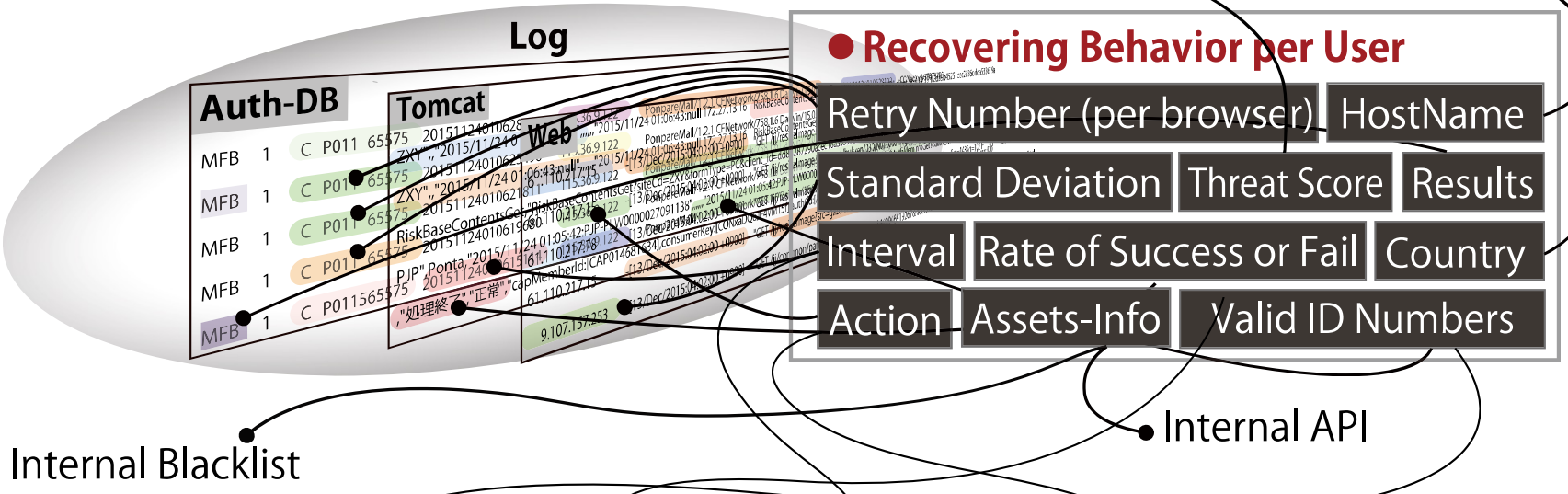


## ①Building BaseQuery (Summary Index)

Whois,TorExitNode,WebAPI,etc

Reverse DNS

Country




Internal API

Internal Blacklist

## ②Heuristic Analysis

AnomaryData + Statistics + Feature + StandardDeviation + Score

**Heuristic Method using Pre and After the Point of Incident**

# Sample Base Query

**Sample BaseQuery (Building 5min ~ Hourly Summary Index** into **index=summary label=base )**

index=Recruit sourcetype=DB_Login_ACT | fillnull UniqID

① |**streamstats** window=0 current=f **last(UniqID) as NextUniqID**
**last(_time) as epochNextTime by SrcIP,BrowserHash** — **Change-Point Detection**

② | fillnull NextUniqID |seach NextUniqID!=0 | **where NextUniqID!=UniqID** — **Threshold**

③ |eval **Interval=epochNextTime-_time** — **Intervals**

④ |stats **count as Total stdev(Interval) as StDev count(eval(AuthCode=1)) as Success**
**min(_time) as min_time max(_time) as max_time by SrcIP,BrowserHash** — **Statistics**

**Another BaseQuery (Building 5min ~ Hourly Summary Index** into **index=summary label=base_si)**

⑥ index=Recruit sourcetype=DB_Login_ACT_APF | fillnull UniqID
|streamstats window=0 current=f last(UniqID) as NextUniqID by SrcIP,BrowserHash
|fillnull NextUniqID |seach NextUniqID!=0 | where NextUniqID!=UniqID → make Macro as `**Anomaly**`

⑤ |**sistats values(UniqID) values(RecruitID) by SrcIP,BrowserHash** — **Statistics & for Merge**

# Sample Heuristic Analysis

**Sample Hourly or Daily Heuristic Analysis from BaseQuery (Using Summary Index )**

**index=summary label=base**

① |stats **count as slot  sum(dc_Success) as sum_dc_Success sum(Total) as sum_Total avg(StDev) as StDev min(min_time) as Start max(max_time) as End by SrcIP,BrowserHash**

② |**JOIN** type=left **SrcIP,UA** [search **index=summary label=base_si**
|stats **dc(UniqID) as ss_dc_UniqID dc(RecruitID) as ss_dc_RecruitID by SrcIP,BrowserHash**
|fields SrcIP,UA, ss_dc_UniqID,ss_dc_RecruitID]

③ |eval **SuccessRate=round(sum_dc_Success/sum_Total*100,2)**
|eval **ValidIDRate=round(ss_dc_RecruitID/ss_dc_UniqID*100,2)**

④ |search ss_dc_RecruitID>=10 | iplocation SrcIP | lookup dnslookup SrcIP

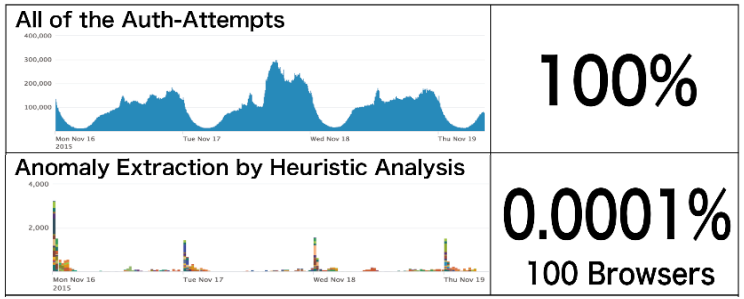
⑤ |table **Start End SrcIP clienthost Country slot *Rate  sum_*  StDev ss_*** | sort-SuccessRate

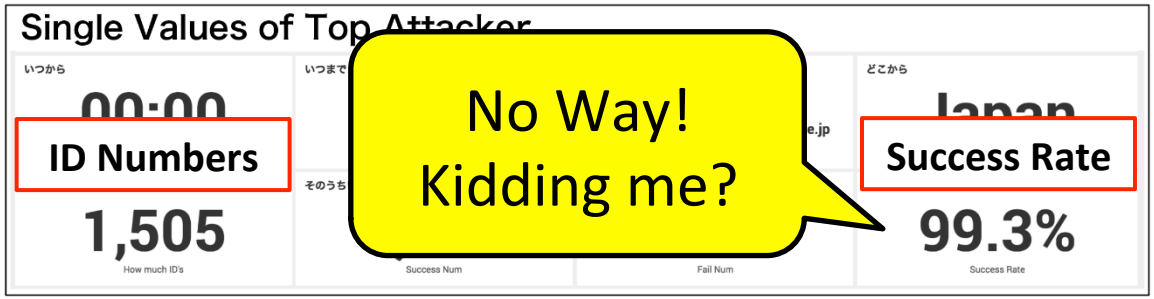
**Feature Extraction**

**More Statistics**

As you like..

**Output**

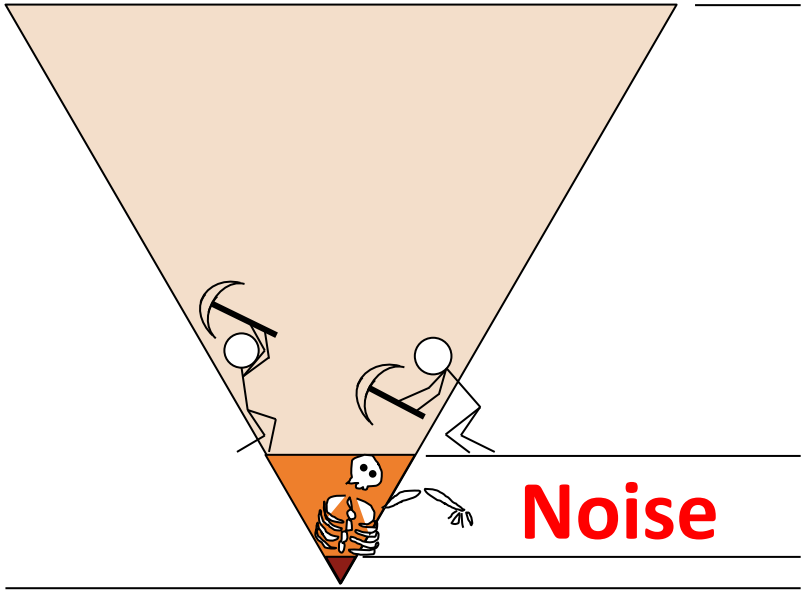


# Splunk Helps Identify Cyber Crimes & Cheats.

## Mystery Data follow-up survey

## Using SPL.

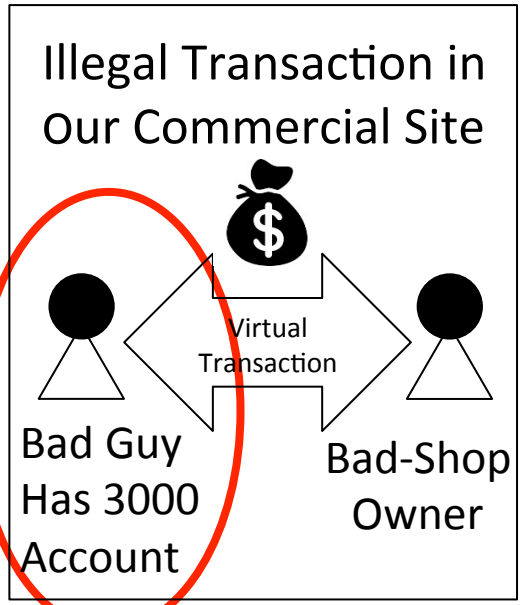



Treasure
(Account hacking)

The One is..

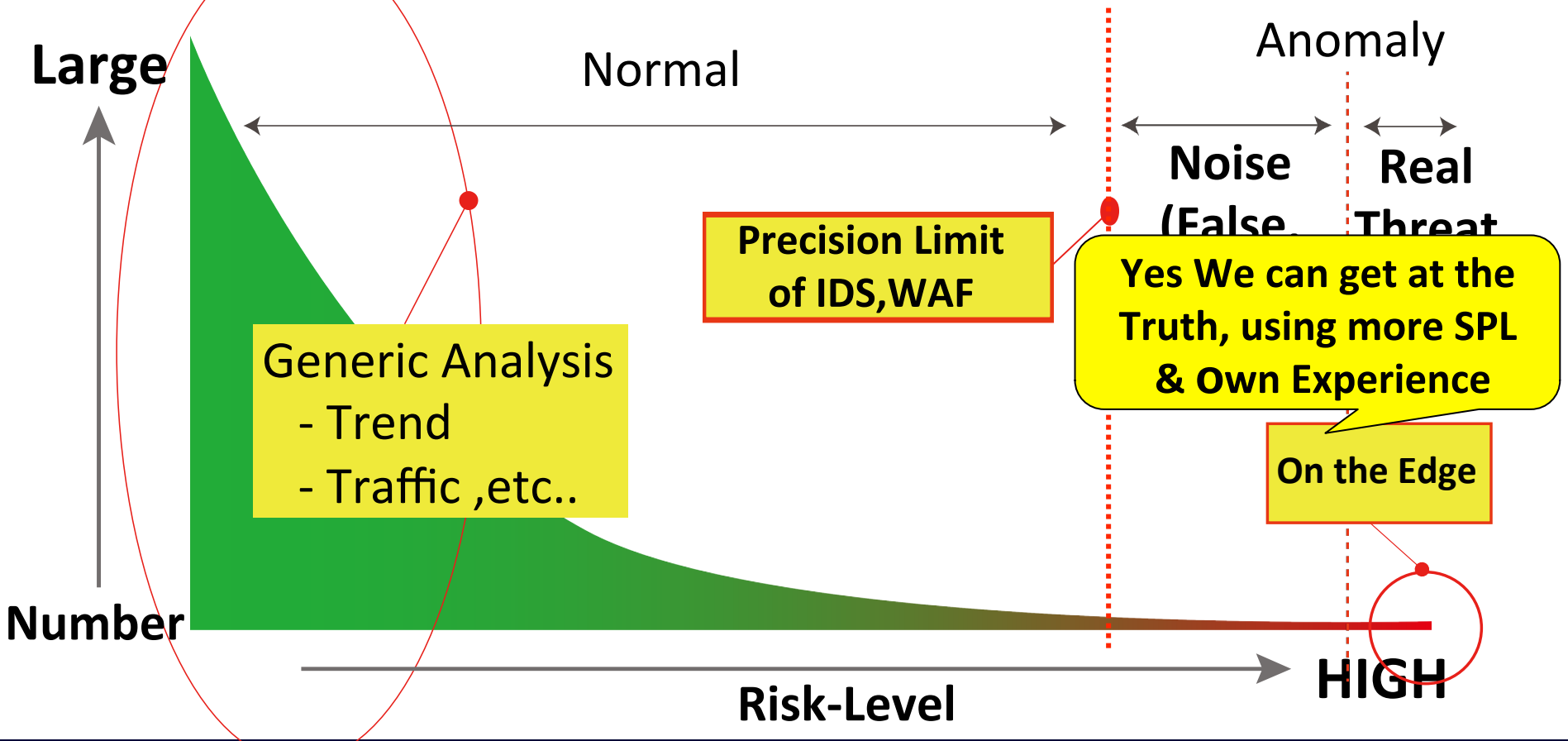

1st Party Frauds

The Other is..




Cyber Crime

# Threats Remain On The Edge Of The Long Tail



# Real-threat Extraction With Customdetectionrule

**Sample Threat Extraction from AnomalyBaseQuery (Using Summary Index )**

① index=summary label=**AnomalyBase NOT** [search index=Recruit **sourcetype=Apache URL="/pointGive/doComp/" [search `anomaly` ]** | streamstats last(_time) as NextTime by SrcIP| eval Interval=NextTime-_time|**search Interval<=10** |fields SrcIP]

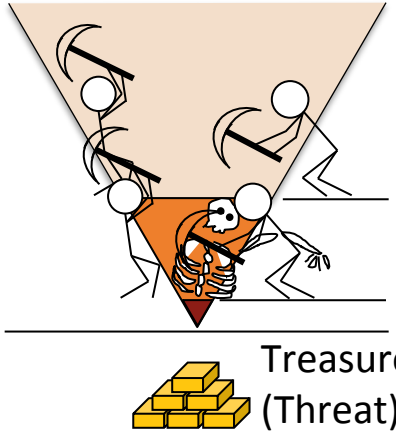
② |**JOIN SrcIP,BrowserHash** [search index=Recruit **sourcetype=Tom_Act [search `anomaly` ]** |stats count(eval(**action="RegID"** AND method="done")) as ct_ID_reg by SrcIP,BrowserHash | fields SrcIP,BrowserHash,ct_ID_reg ]

③ **| search StDev<=0.9 ct_ID_reg>=10 | timechart span=10m count by SrcIP**

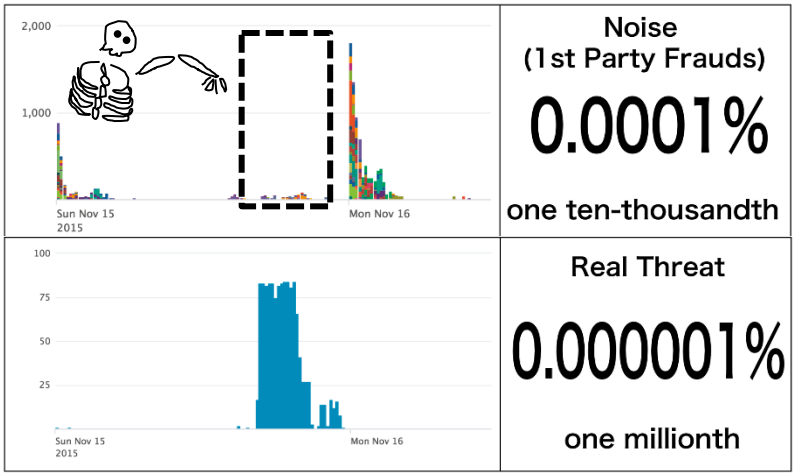
**Sift out NOISE using Custom-Rule, and throw away**

**Feature Extraction for specify Threat**

**Final Output**



Treasure (Threat)



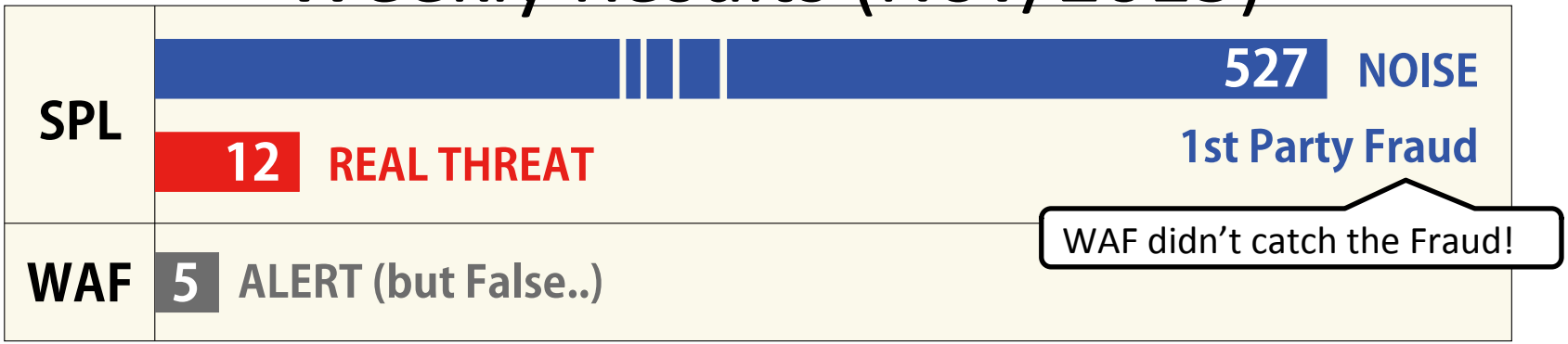

**Registrations of 5000+ IDs by one person.**

Another Observed Threat.

- **List-Type Account Hacking.**
- **The Act of removing Dummy ID's.**
- **Online Virtual-Coin Laundering.**

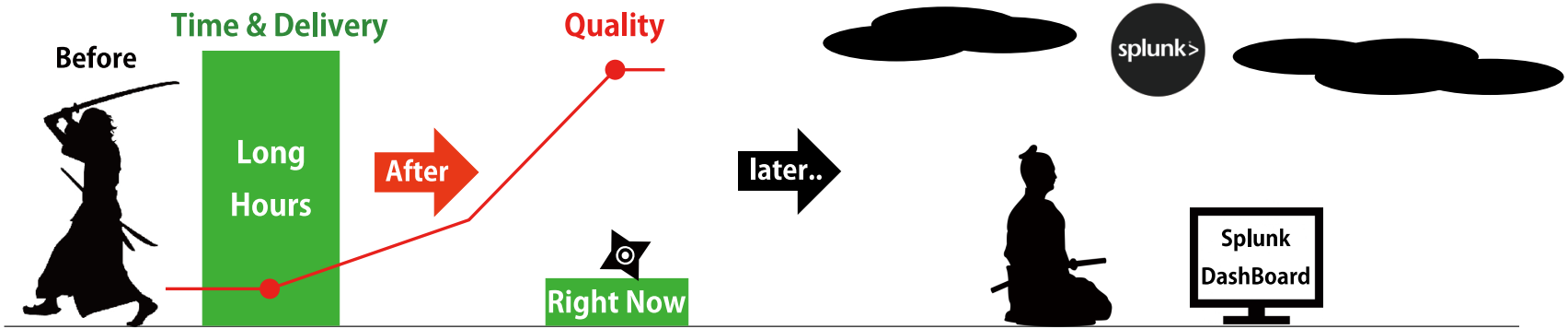
**etc…**

# Weekly Results (Nov/2015)



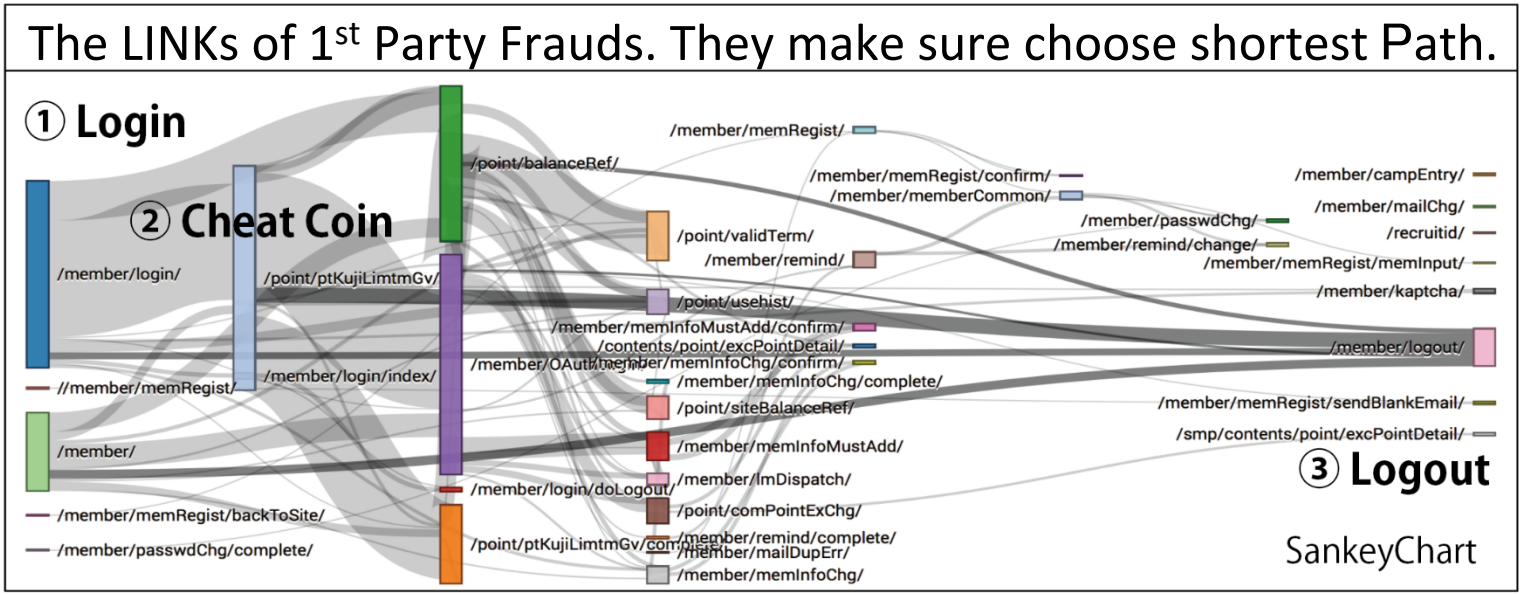

**Their DeepAnalysis has changed dramatically.**

**They can judge things calmly..**



# Challenge 2:
# Data Visualization Using d3.js

# Data-Driven Documents

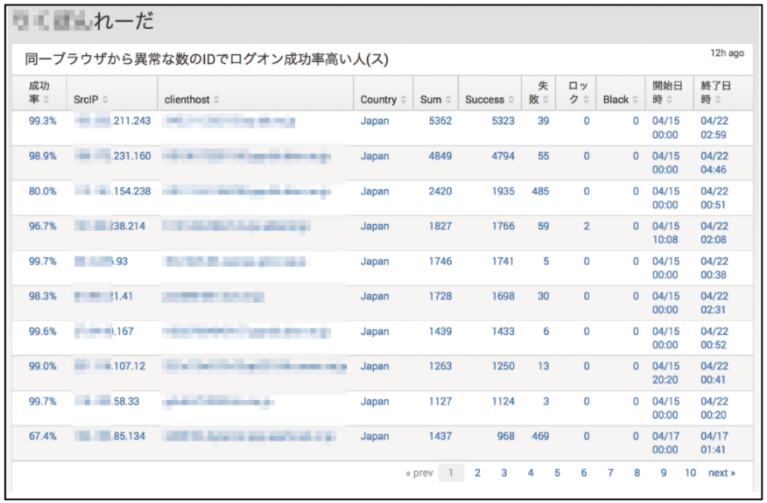



**D3.js** is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

See more examples.

https://d3js.org/

# Data-visualization Of 1st Party Frauds (Noise)



れーだ

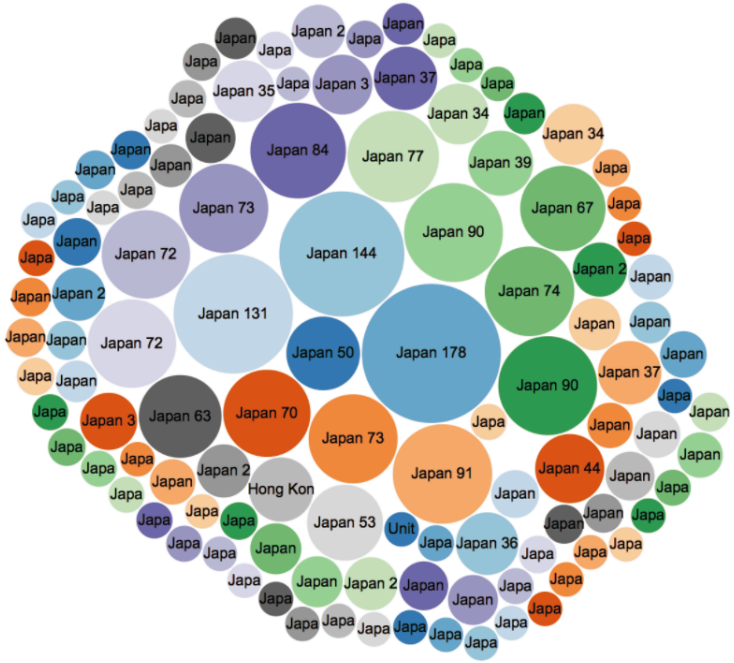
同一ブラウザから異常な数のIDでログオン成功率高い人(ス)

12h ago

| 成功率 | SrcIP | clienthost | Country | Sum | Success | 失敗 | ロック | Black | 開始日時 | 終了日時 |
|---|---|---|---|---|---|---|---|---|---|---|
| 99.3% | 211.243 | | Japan | 5362 | 5323 | 39 | 0 | 0 | 04/15 00:00 | 04/22 02:59 |
| 98.9% | 231.160 | | Japan | 4849 | 4794 | 55 | 0 | 0 | 04/15 00:00 | 04/22 04:46 |
| 80.0% | 154.238 | | Japan | 2420 | 1935 | 485 | 0 | 0 | 04/15 00:00 | 04/22 00:51 |
| 96.7% | 38.214 | | Japan | 1827 | 1766 | 59 | 2 | 0 | 04/15 10:08 | 04/22 02:08 |
| 99.7% | 5.93 | | Japan | 1746 | 1741 | 5 | 0 | 0 | 04/15 00:00 | 04/22 00:38 |
| 98.3% | 1.41 | | Japan | 1728 | 1698 | 30 | 0 | 0 | 04/15 00:00 | 04/22 02:31 |
| 99.6% | .167 | | Japan | 1439 | 1433 | 6 | 0 | 0 | 04/15 00:00 | 04/22 00:52 |
| 99.0% | 107.12 | | Japan | 1263 | 1250 | 13 | 0 | 0 | 04/15 20:20 | 04/22 00:41 |
| 99.7% | 58.33 | | Japan | 1127 | 1124 | 3 | 0 | 0 | 04/15 00:00 | 04/22 00:20 |
| 67.4% | 85.134 | | Japan | 1437 | 968 | 469 | 0 | 0 | 04/17 00:00 | 04/17 01:41 |

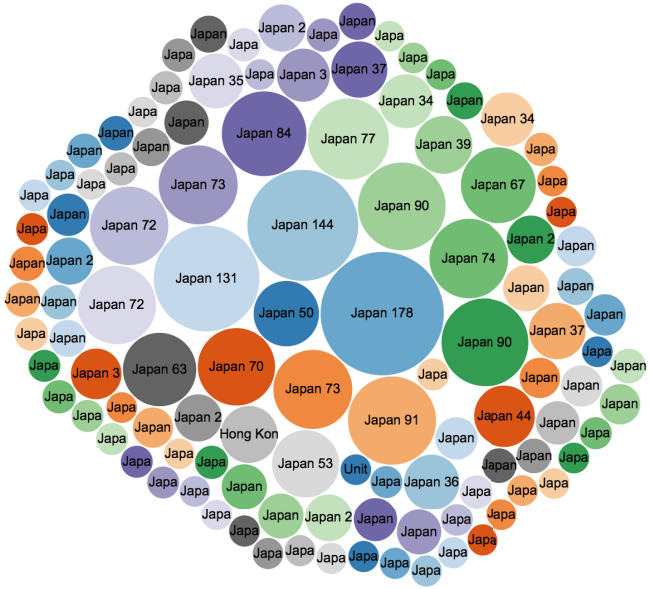
« prev 1 2 3 4 5 6 7 8 9 10 next »

Table

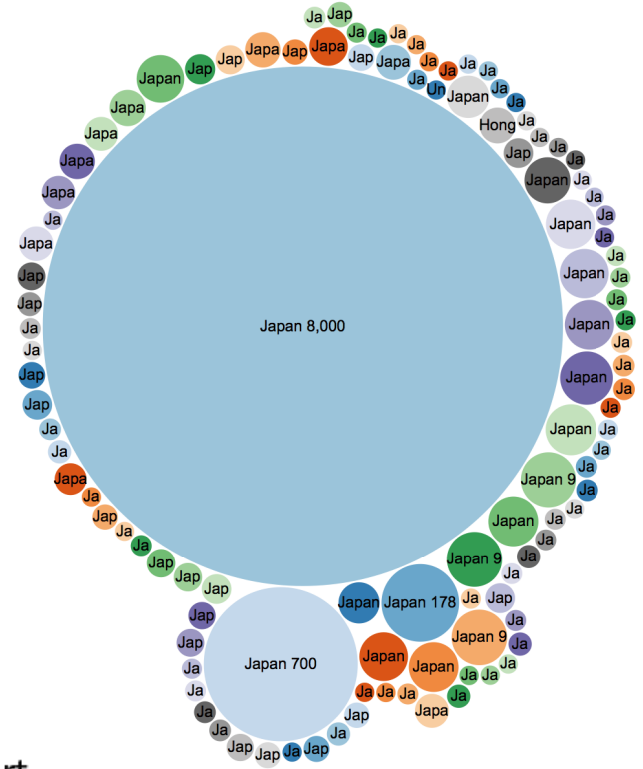



BubbleChart

# Normal (Including Noise)




# List-Type Account Hacking




BubbleChart

Visible All from the Above

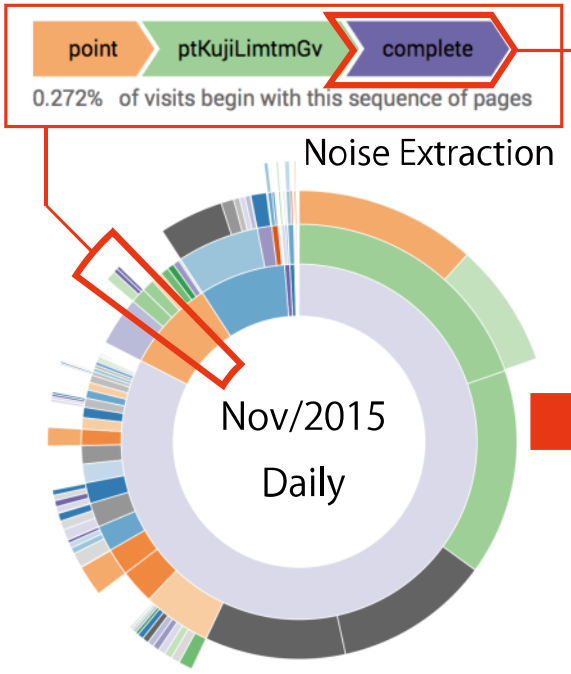
**100%**

Noise Extraction

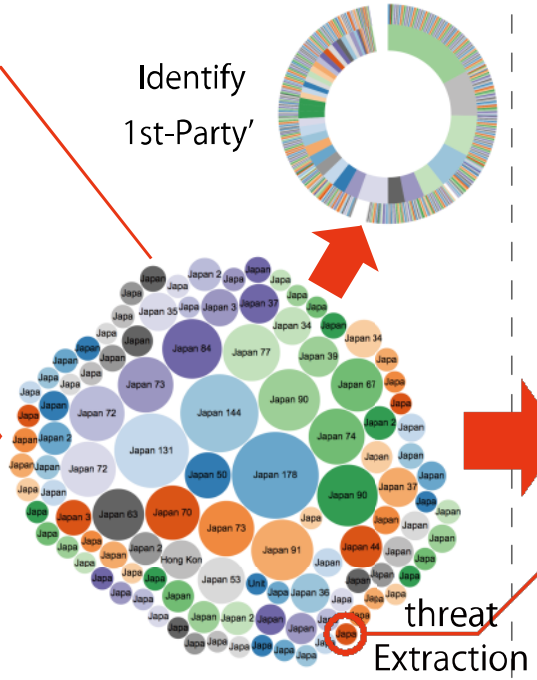
**0.0001%**

Real Threat

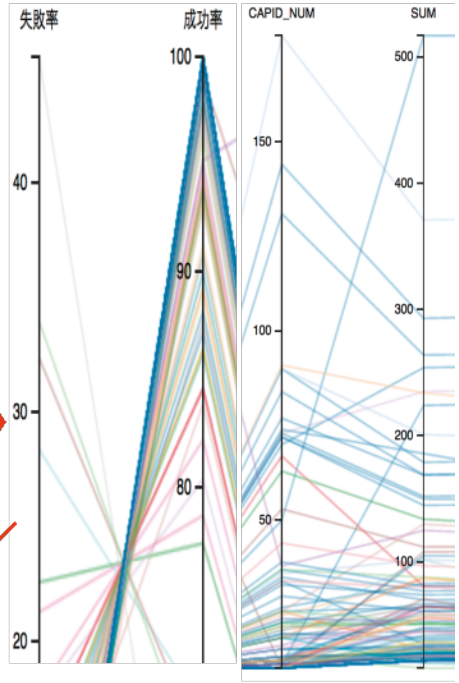
**0.000001%**




Sunburst



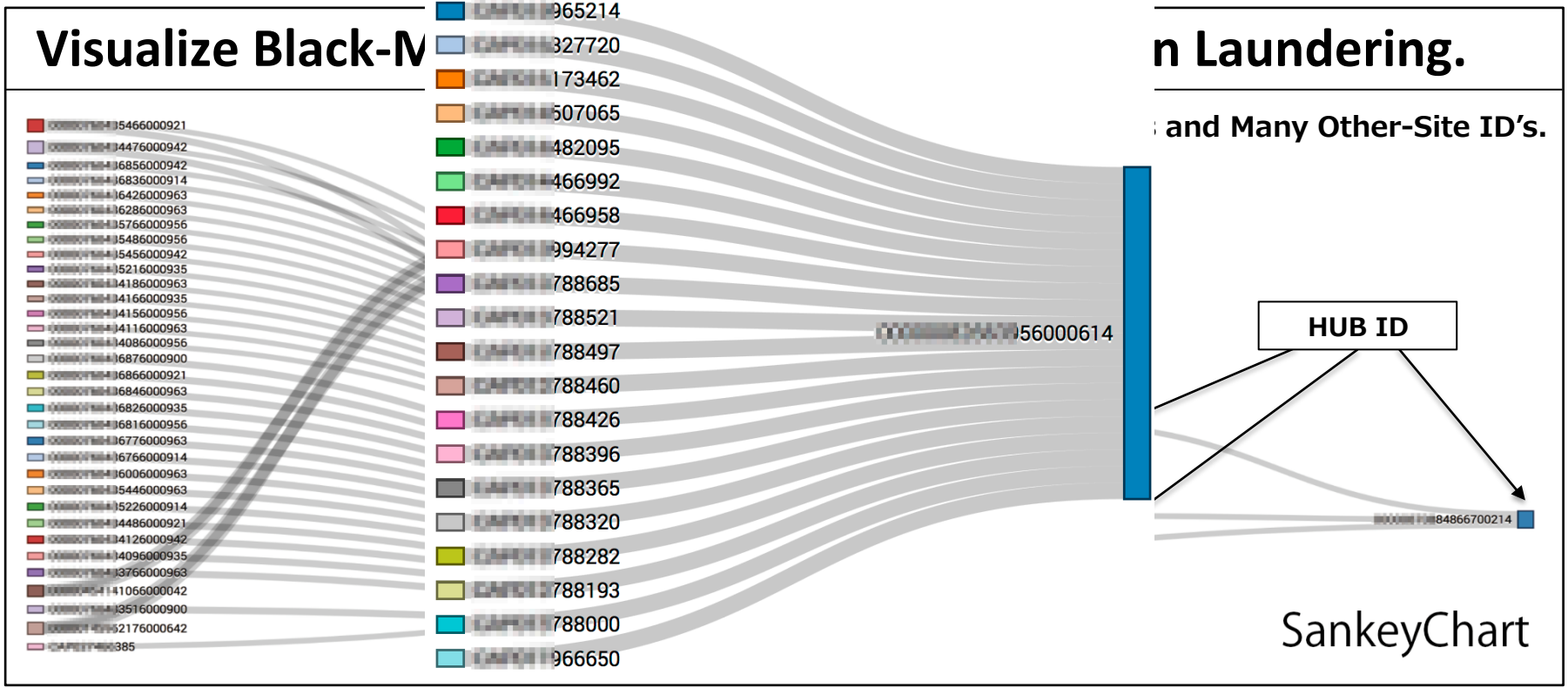

BubbleChart




Parallel Coordinates

# Case Studies

# Identify and Visualize underlying Threat



# Two Dashboards

## For extract Threat-Seeds

## For extract Real Threat


## For Future IR

## For Rapid IR

# Wrap Up

# Wrap Up

1. Recovering Behavior from Fragment Data.

   Using SummaryIndex, using Stats, StreamStats, and Sistats

2. Classify Anomaly Data as to NOISE and THREAT by Heuristic Analysis.

   Extraction Accurancy had reached a millionth

3. Stuff who responding Incident in Recruit make rapid progress.

   Splunk defeat WAF in the Recruit!!

4. Splunk is not only for Rapid-IR,  but also for Future-IR

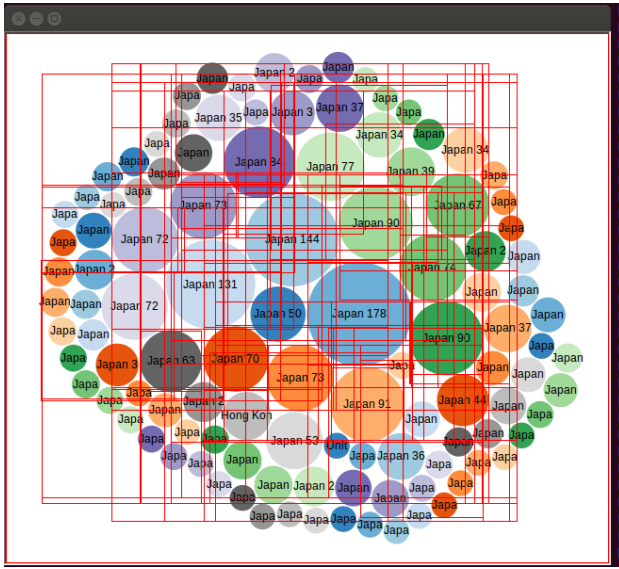
5. Data-Visualization has infinite possibilities

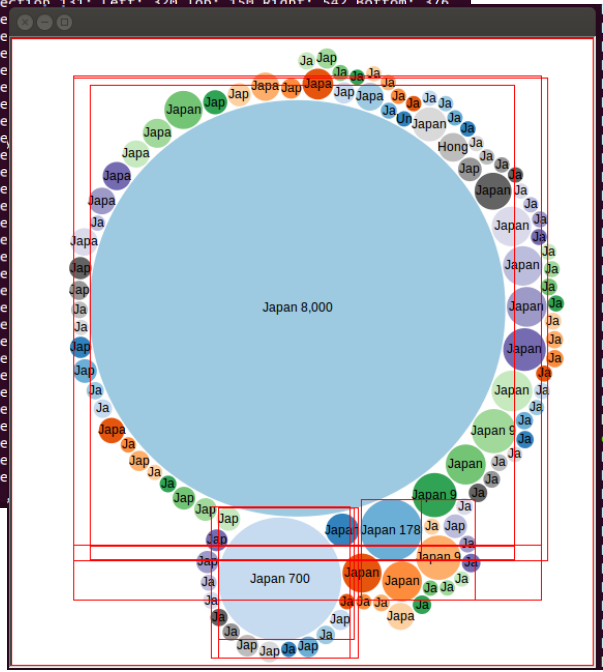
# Extra Challenge
# Deep Learning

## d3.js-Image Recognition

# Threat Detection With Deeplearning (Without Training-data)

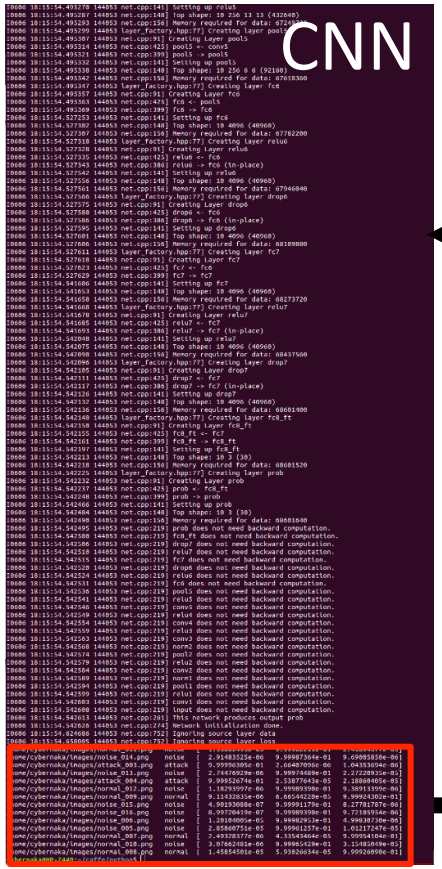
## Object Recognition using “Dlib” & ” Scikit-Image”






```
Detection 128: Left: 280 Top: 61 Right: 497 Bottom: 240
Detection 129: Left: 281 Top: 55 Right: 421 Bottom: 207
Detection 130: Left: 281 Top: 55 Right: 468 Bottom: 230
Detection 131: Left: 320 Top: 150 Right: 542 Bottom: 376
Detection 138: Left: 331 Top: 310 Right: 431 Bottom: 432
Detection 139: Left: 339 Top: 96 Right: 497 Bottom: 240
Detection 140: Left: 342 Top: 313 Right: 443 Bottom: 432
Detection 141: Left: 342 Top: 313 Right: 506 Bottom: 493
Detection 142: Left: 345 Top: 285 Right: 505 Bottom: 434
Detection 143: Left: 345 Top: 313 Right: 475 Bottom: 434
Detection 144: Left: 345 Top: 355 Right: 422 Bottom: 432
Detection 145: Left: 354 Top: 150 Right: 542 Bottom: 252
Detection 146: Left: 354 Top: 150 Right: 542 Bottom: 283
Detection 147: Left: 354 Top: 162 Right: 462 Bottom: 240
Detection 148: Left: 354 Top: 163 Right: 430 Bottom: 240
Detection 149: Left: 369 Top: 313 Right: 475 Bottom: 434
Detection 150: Left: 376 Top: 331 Right: 506 Bottom: 514
Detection 151: Left: 376 Top: 347 Right: 506 Bottom: 493
Detection 152: Left: 418 Top: 150 Right: 542 Bottom: 283
Detection 153: Left: 418 Top: 204 Right: 505 Bottom: 283
Detection 154: Left: 418 Top: 214 Right: 487 Bottom: 283
Detection 155: Left: 428 Top: 285 Right: 505 Bottom: 362
Detection 156: Left: 428 Top: 285 Right: 534 Bottom: 372
Detection 157: Left: 429 Top: 285 Right: 534 Bottom: 372
Detection 158: Left: 446 Top: 150 Right: 512 Bottom: 216
Detection 159: Left: 446 Top: 150 Right: 542 Bottom: 252
Hit enter to continue
cybernaka@HP-Z440:~/caffe/python$ python dlib_classify.py list.png
Detection 0: Left: 0 Top: 0 Right: 562 Bottom: 605
Detection 1: Left: 60 Top: 36 Right: 512 Bottom: 489
Detection 2: Left: 60 Top: 36 Right: 512 Bottom: 542
Detection 3: Left: 60 Top: 38 Right: 518 Bottom: 504
Detection 4: Left: 76 Top: 45 Right: 486 Bottom: 490
Detection 5: Left: 76 Top: 45 Right: 486 Bottom: 503
Detection 6: Left: 193 Top: 452 Right: 327 Bottom: 598
Detection 7: Left: 200 Top: 453 Right: 331 Bottom: 580
Detection 8: Left: 200 Top: 453 Right: 335 Bottom: 598
Detection 9: Left: 338 Top: 445 Right: 396 Bottom: 503
Detection 10: Left: 338 Top: 445 Right: 448 Bottom: 542
Hit enter to continue
```

# Threat Detection With Deeplearning (With Training-data)



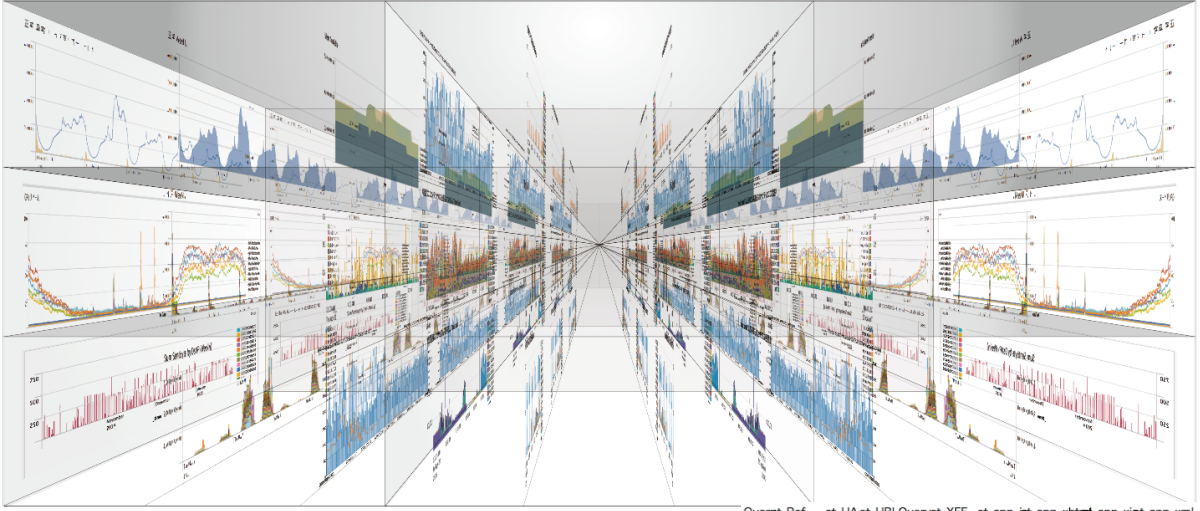

**Input : d3.js Images**


▽**Technologies**
- **CUDA**
- **Caffe**
- **SelectiveSearch**
- **PyData, etc..**

▽**Who gains benefit ?**
**24/7 SOC**

**Output : Whether Man or Machine, Anyone can see Attack-Image is Threat !**

```
/home/cybernaka/images/normal_011.png    noise    [  1.35602495e-05   9.99962211e-01   2.41894377e-05]
/home/cybernaka/ima[illegible]           n[illegible]  [  2.91483525e-06   9.99987364e-01   9.69005850e-06]
/home/cybernaka/images/attack_003.png    attack   [  9.99996305e-01   2.66407096e-06   1.04353694e-06]
/home/cybernaka/ima[illegible]/noise_013.png    noise    [  2.74476929e-06   9.99974489e-01   2.27220935e-05]
/home/cybernaka/ima[illegible]/attack_004.png   attack   [  9.99952674e-01   2.53877643e-05   2.18860405e-05]
/home/cybernaka/imag[illegible]          n[illegible]  [  1.18293997e-06   9.99989390e-01   9.38913399e-06]
/home/cybernaka/images/normal_009.png    normal   [  9.11432835e-06   6.66544220e-05   9.99924302e-01]
/home/cybernaka/images/noise_015.png     noise    [  4.90193088e-07   9.99991179e-01   8.27781787e-06]
/home/cybernaka/images/noise_016.png     noise    [  8.99720419e-07   9.99989390e-01   9.72105954e-06]
/home/cybernaka/images/noise_006.png     noise    [  1.20104005e-05   9.99982953e-01   4.99830730e-06]
/home/cybernaka/images/noise_005.png     noise    [  2.85860751e-05   9.99961257e-01   1.01217247e-05]
/home/cybernaka/images/normal_007.png    normal   [  2.49328377e-06   4.33543464e-05   9.99954104e-01]
/home/cybernaka/images/normal_010.png    noise    [  3.07662481e-06   9.99965429e-01   3.15485049e-05]
/home/cybernaka/images/normal_008.png    normal   [  1.45854501e-05   5.93026634e-05   9.99926090e-01]
cybernaka@HP-Z440:~/caffe/python$
```

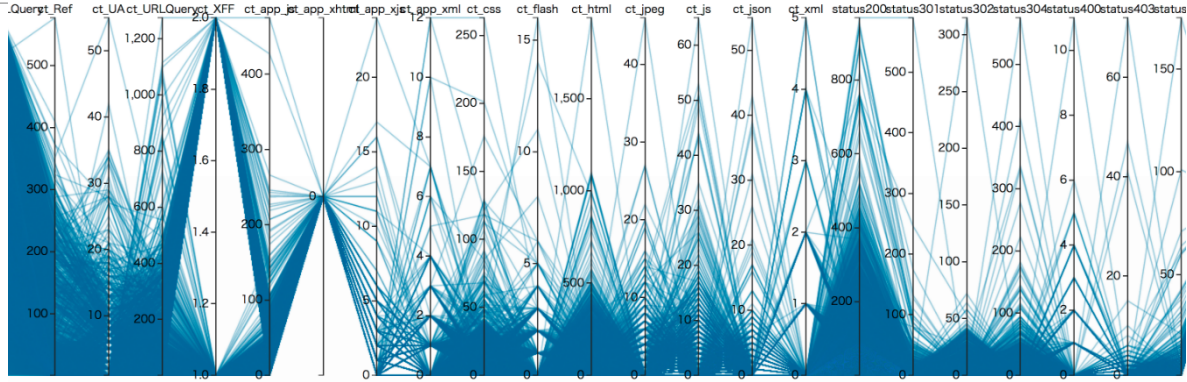


Experts & Long Time

↓

Deep Learning
and Training-Data

Data-Visualization




**Parallel Coordinates**

**A visual toolkit for multidimensional detectives.**

**https://github.com/syntagmatic/parallel-coordinates**

| TotalDur... | HttpMet... | TOP_UR... | ct_Ref | ct_UA | ct_URLQ... | ct_XFF | ct_app_j... | ct_app_... | ct_app_... | ct_app_... | ct_css | ct_flash | ct_html | ct_jpeg | ct_js | ct_json | ct_xml | status20... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 159 | GET | / | 1 | 1 | 68 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 68 | 0 | 0 | 0 | 0 | 1 |
| 506 | GET HEAD | /secure/... | 1 | 1 | 23 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 1 |
| -7326 | GET | /wp-con... | 1 | 1 | 39 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 44 | 0 | 0 | 0 | 0 | 2 |

# THANK YOU

Special Thanks: Mifune-san, Splunk-Japan, and Macnica

hiro:<nakamura@r.recruit.co.jp>

Facebook : https://www.facebook.com/hiro.goahead