Background

Observation

Traffic Analysis - Cluster by daily volume

DS Team

Implement

Thinking

First of All

Anti-Spam Solution

Anti-Spam Service

Check the reputation of sender IP

Blocked if the reputation is BAD

Receive

Check the reputation of the email hash”s" of sender IP

Blocked if the email hash”s” is BAD

Receive

a. Sender IP

b. Email Hash”s”

Background

$$\text{IP Reputation} = \frac{\text{Spam Volume}}{\text{Request Volume}}$$

Background

$$\text{IP Reputation} = \frac{\text{Spam Volume}}{\text{Request Volume}}$$

Good $\sim \dfrac{\text{Spam Volume}}{\text{Request Volume}}$

Bad $\sim \dfrac{\text{Spam Volume}}{\text{Request Volume}}$

Background

冠穎什麼時候不再罵自己笨

哪一類型的 IP 信譽是調動不斷的, 或是哪個時間點

300~400M requests

一天中哪個時段信件最多

垃圾郵件最多是從哪個網斷劑出來，哪個國家最多

1 day from customers

我們發生 False Positive 的郵件類型是哪一種

采襄什麼時候可以不用做簡報

40~50M spam emails

Roger 什麼時候跟姿吟講話？

哪一類型的垃圾郵件最多

哪個客戶的信件量越來越少

Miles已經會用 vim 開發程式了

1 day from spam catcher

哪種語言我們做的最差

客戶是不是有週期性的症狀

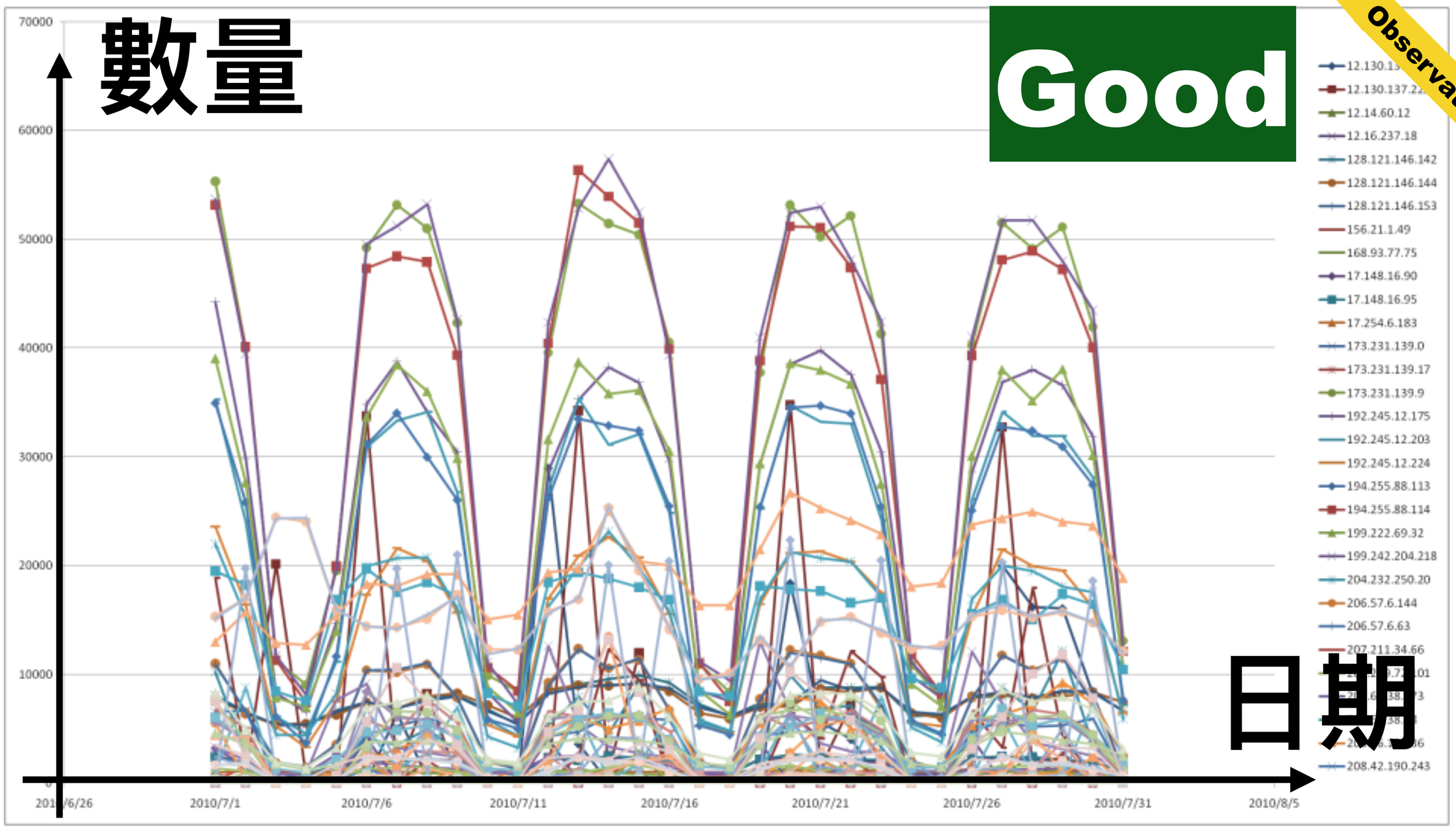
Background

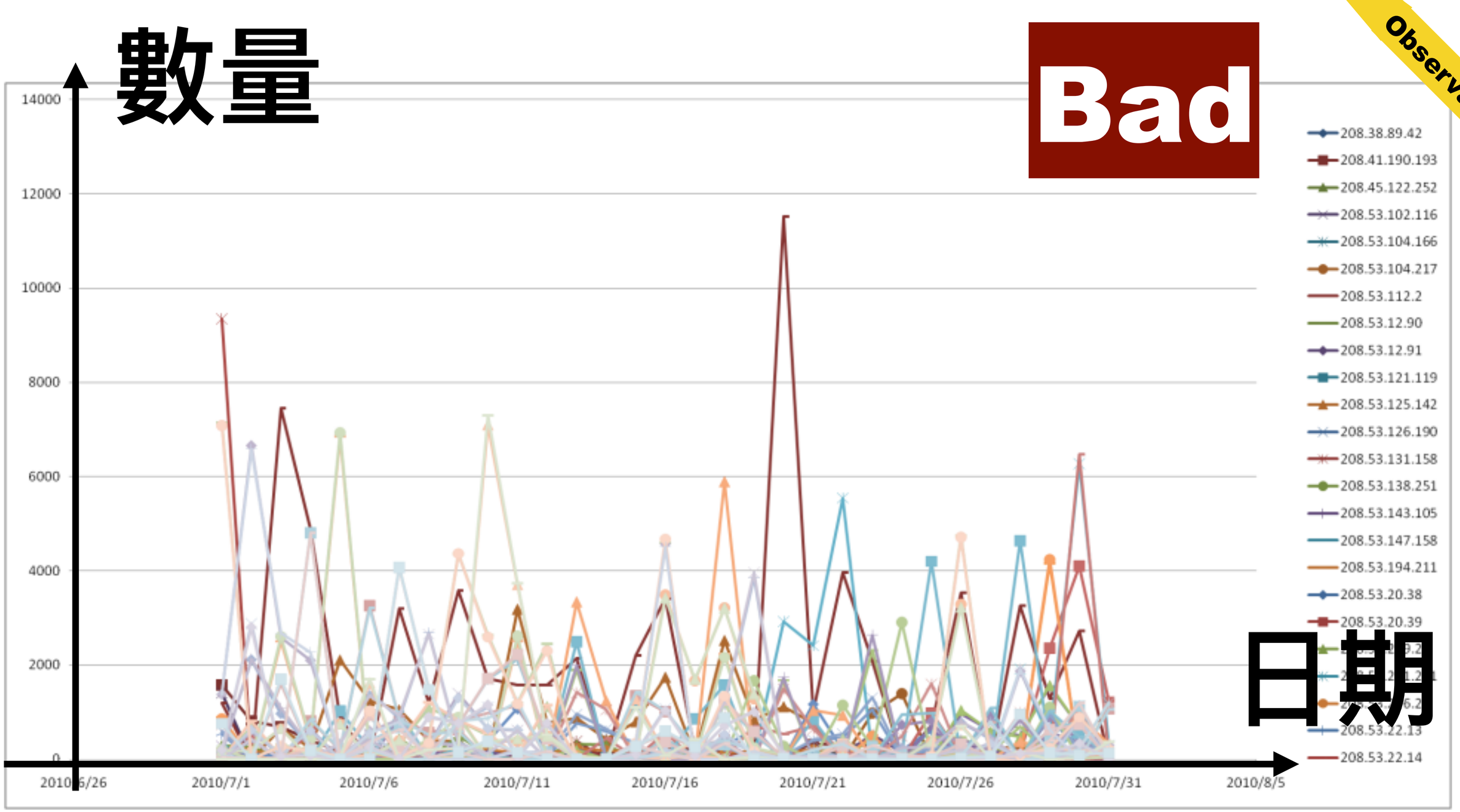
About 1.5 years

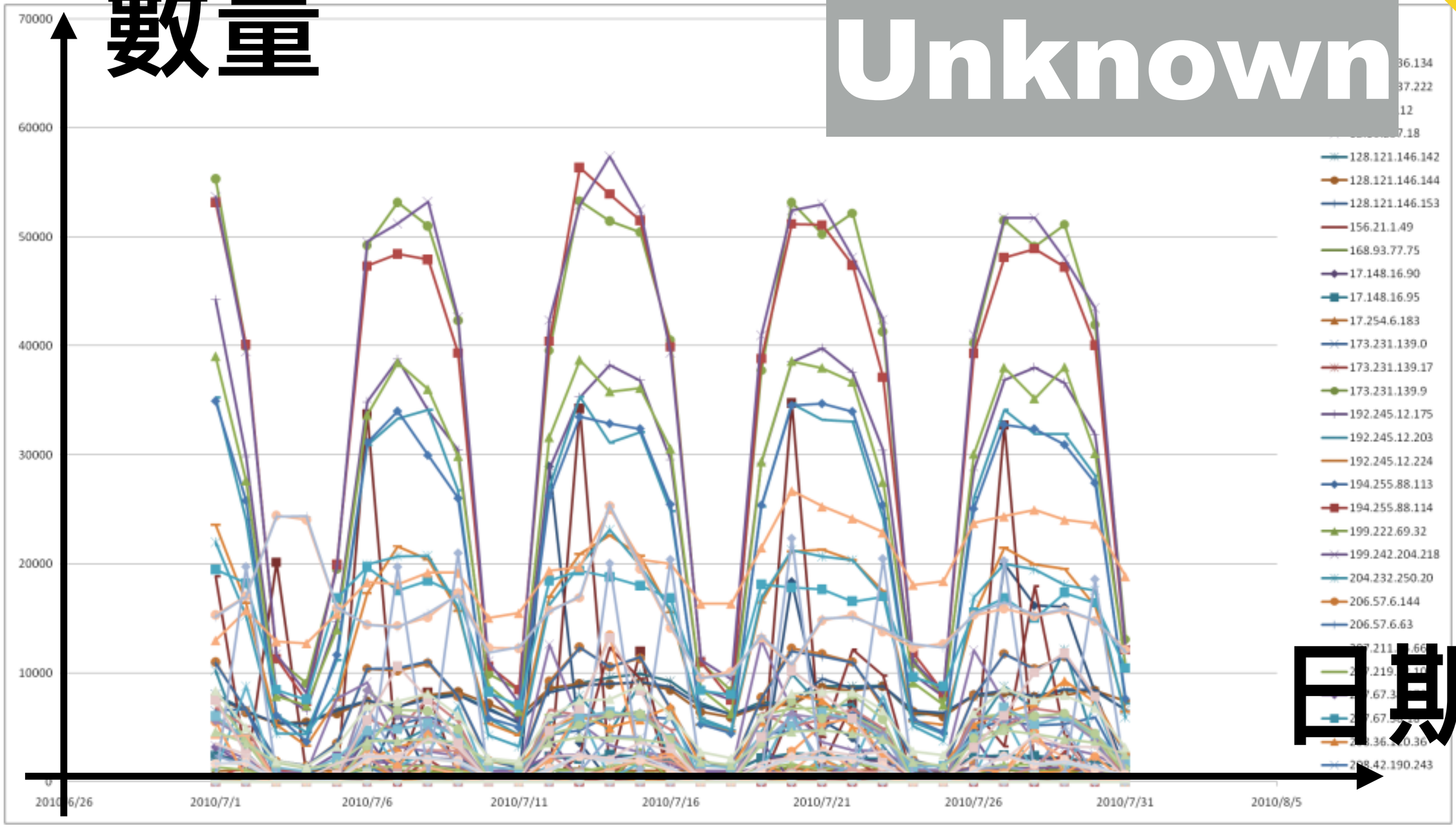
1. Develop programs and develop myself engineering skills
2. Know your number (Like 己茵)
3. Understand / Review current concept of IP Reputation formula

$$\text{IP Reputation} = \frac{\text{Spam Volume}}{\text{Request Volume}}$$

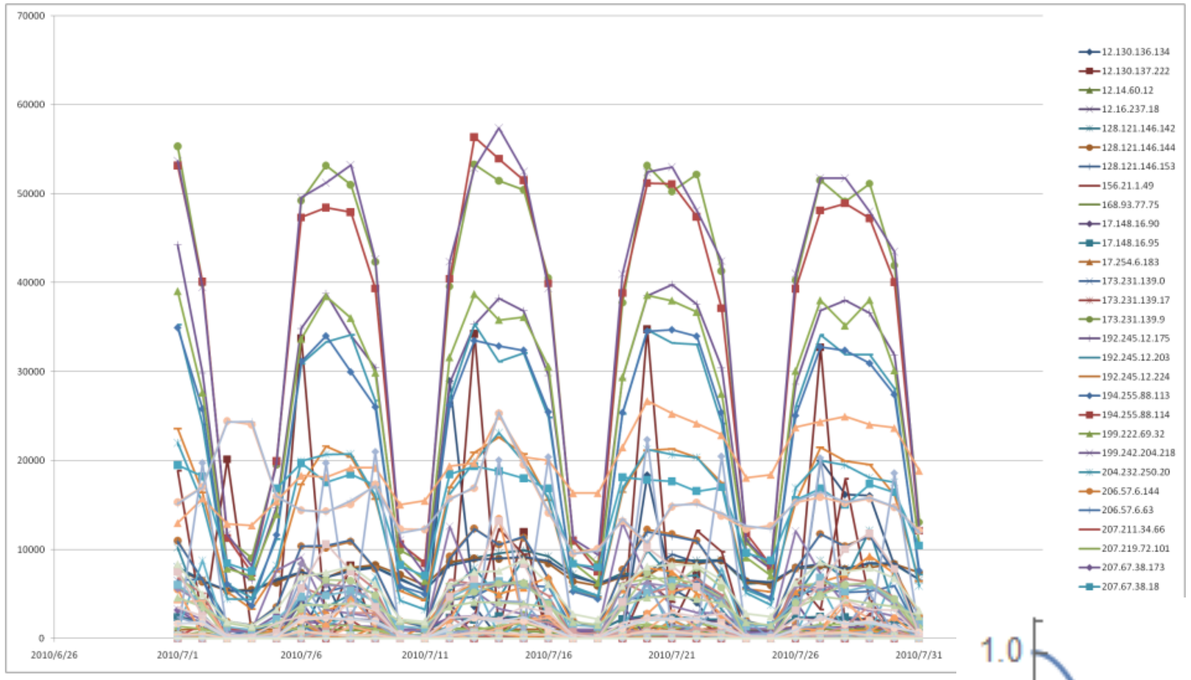
If we can collect the whole email traffic, this formula is wonderful!!!


Observation
Good
数量
日期
70000
60000
50000
40000
30000
20000
10000
0
2010/6/26
2010/7/1
2010/7/6
2010/7/11
2010/7/16
2010/7/21
2010/7/26
2010/7/31
2010/8/5
12.130.13
12.130.137.22
12.14.60.12
12.16.237.18
128.121.146.142
128.121.146.144
128.121.146.153
156.21.1.49
168.93.77.75
17.148.16.90
17.148.16.95
17.254.6.183
173.231.139.0
173.231.139.17
173.231.139.9
192.245.12.175
192.245.12.203
192.245.12.224
194.255.88.113
194.255.88.114
199.222.69.32
199.242.204.218
204.232.250.20
206.57.6.144
206.57.6.63
207.211.34.66
208.42.190.243


Observation
Bad
數量
日期
14000
12000
10000
8000
6000
4000
2000
0
2010/6/26
2010/7/1
2010/7/6
2010/7/11
2010/7/16
2010/7/21
2010/7/26
2010/7/31
2010/8/5
208.38.89.42
208.41.190.193
208.45.122.252
208.53.102.116
208.53.104.166
208.53.104.217
208.53.112.2
208.53.12.90
208.53.12.91
208.53.121.119
208.53.125.142
208.53.126.190
208.53.131.158
208.53.138.251
208.53.143.105
208.53.147.158
208.53.194.211
208.53.20.38
208.53.20.39
208.53.22.13
208.53.22.14

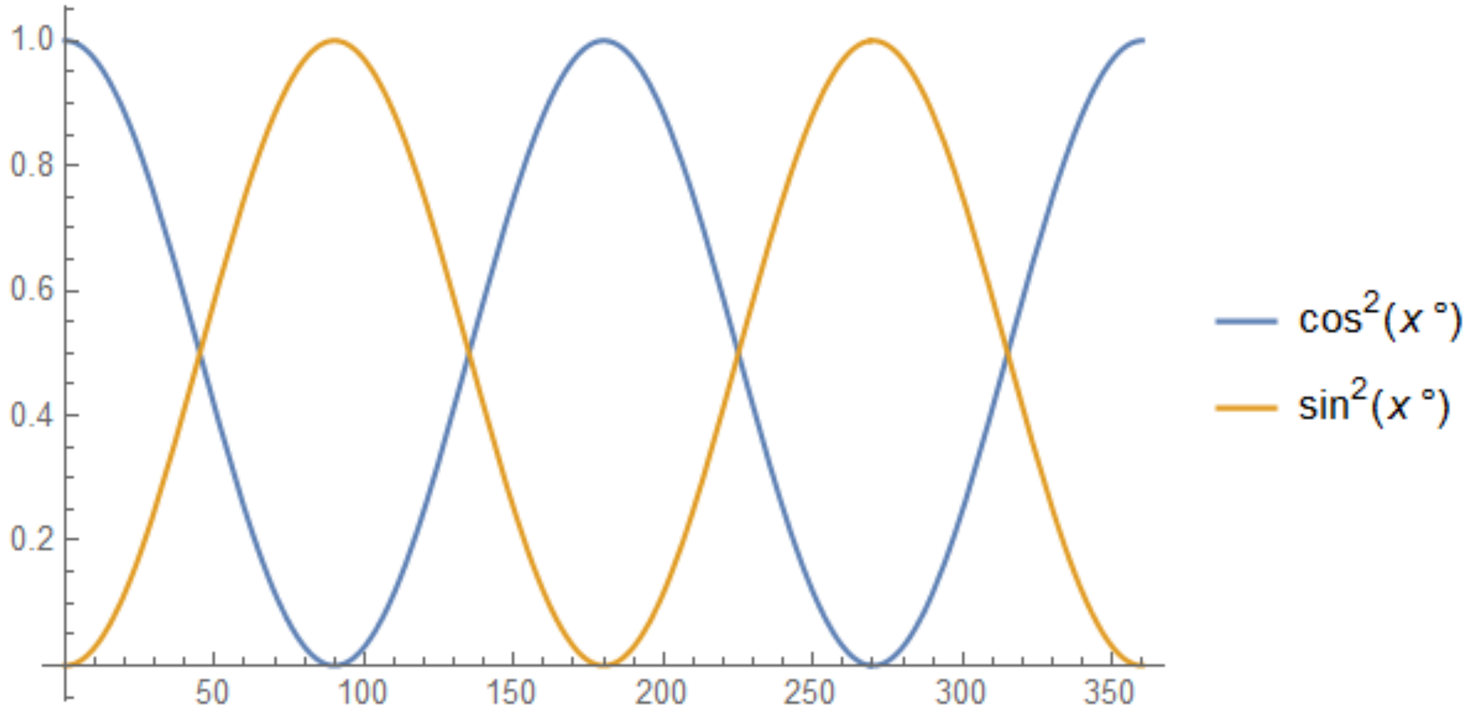

数量
Unknown
Observation
70000
60000
50000
40000
30000
20000
10000
0
2010/6/26
2010/7/1
2010/7/6
2010/7/11
2010/7/16
2010/7/21
2010/7/26
2010/7/31
2010/8/5
128.121.146.142
128.121.146.144
128.121.146.153
156.21.1.49
168.93.77.75
17.148.16.90
17.148.16.95
17.254.6.183
173.231.139.0
173.231.139.17
173.231.139.9
192.245.12.175
192.245.12.203
192.245.12.224
194.255.88.113
194.255.88.114
199.222.69.32
199.242.204.218
204.232.250.20
206.57.6.144
206.57.6.63
日期

Observation



- 針對每個IP線條，使用數學公式描述

- 於是腦中浮現了...



於是，惡補了三角函數1W

而得到的結論是...

考不上數學系

不是沒有道理的...

Thinking

但是，金庸連城訣

倒是給了靈感...

Thinking

主角：石破天

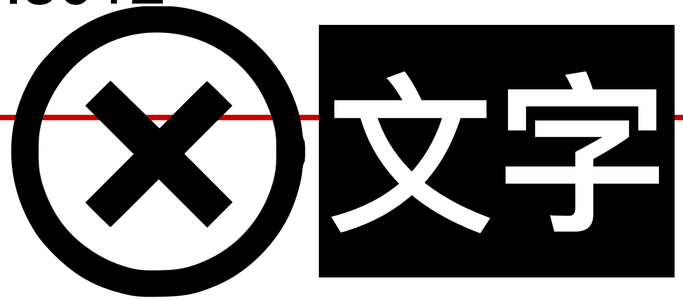
智慧：不識字

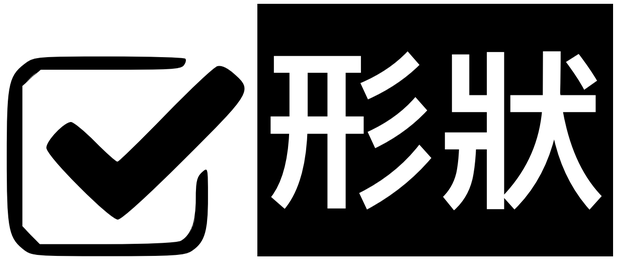
武功：因為不識字所以把壁上的字當圖看，反而練成絕世武功，成為武林第一人...

Thinking

Cialis20mg + 4 FREEvira Pills. No Rx

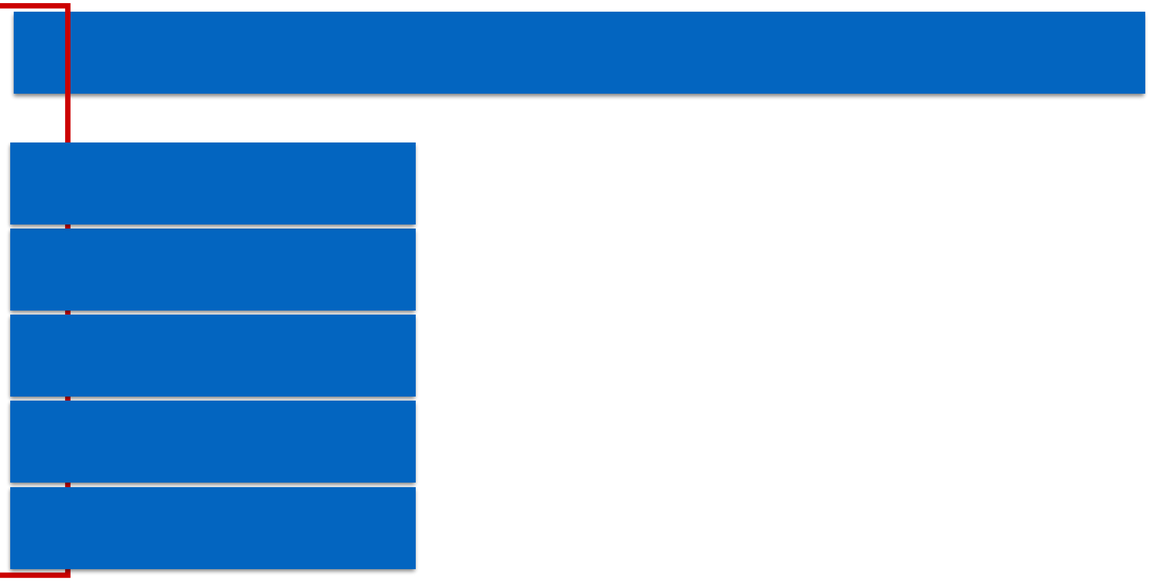
10 pills $52.5
30 pills $136.67
60 pills $238.08
90 pills $310.3012
1 pills $3.08





Cialis31mg + 7 FREEviaqra Pills. No Rx

10 pills $15.21
20 pills $150.07
70 pills $251.09
80 pills $315.21
90 pills $335.08



Implement

Q.

比較兩文件
相似度

Cialis20mg + 4 FREEvira Pills. No Rx

10 pills $52.5
30 pills $136.67
60 pills $238.08
90 pills $310.3012
1 pills $3.08

38 characters	34
New Line	0
14 characters	13
16 characters	13
16 characters	13
18 characters	13
13 characters	13

Q. 字數越少越精確，字數越多越模糊

34	40 characters
0	New Line
13	16 characters
13	15 characters
13	16 characters
13	16 characters
13	17 characters

Cialis31mg + 7 FREEviaqra Pills. No Rx

10 pills $15.21
20 pills $150.07
70 pills $251.09
80 pills $315.21
130 pills $335.08

費氏數列

Implement

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 69 ...

Cialis20mg + 4 FREEviagra Pills. No Rx

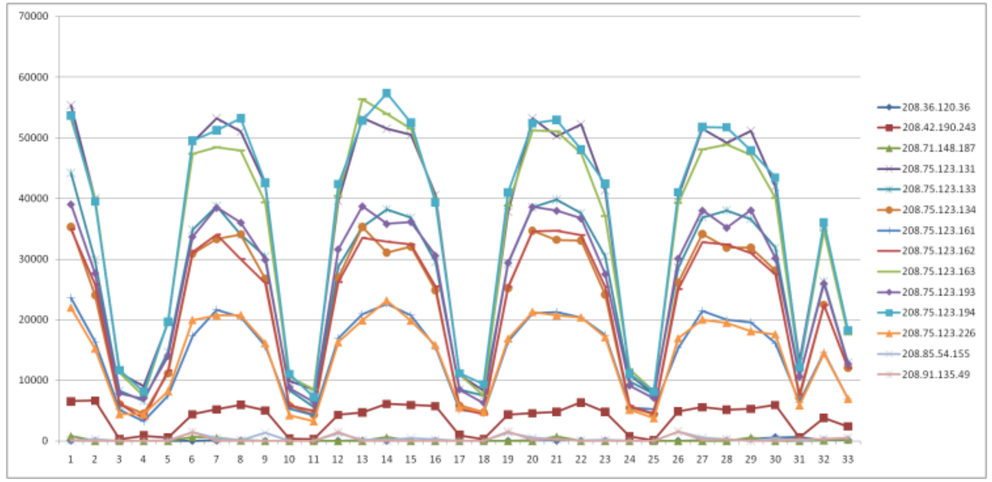
10 pills $52.57
30 pills $136.67
60 pills $238.08
90 pills $310.30
120 pills $334.08

1

Cialis20mg + 4 FREEviagra Pills. No Rx

10 pills $52.57
30 pills $136.67
60 pills $238.08
90 pills $310.30
120 pills $334.08

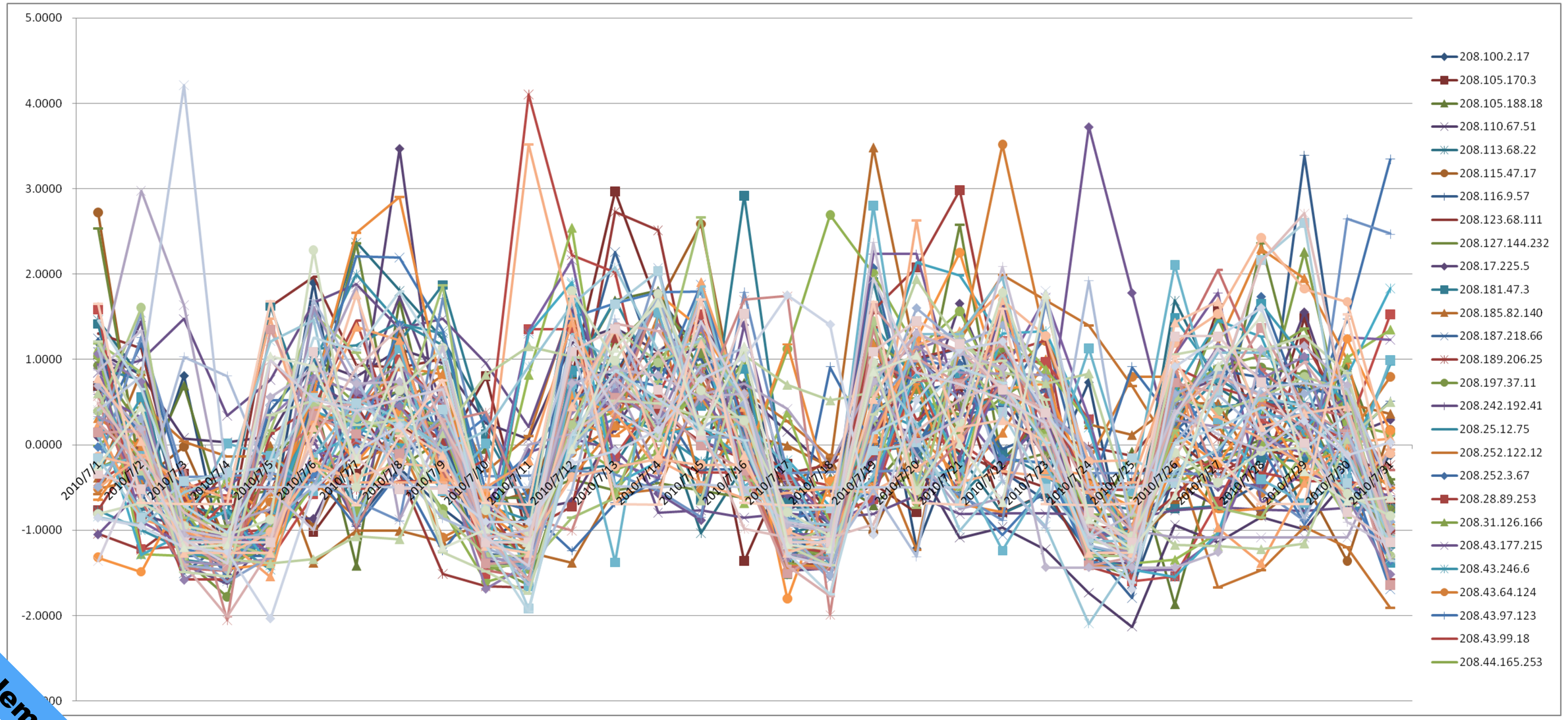
2



Implement

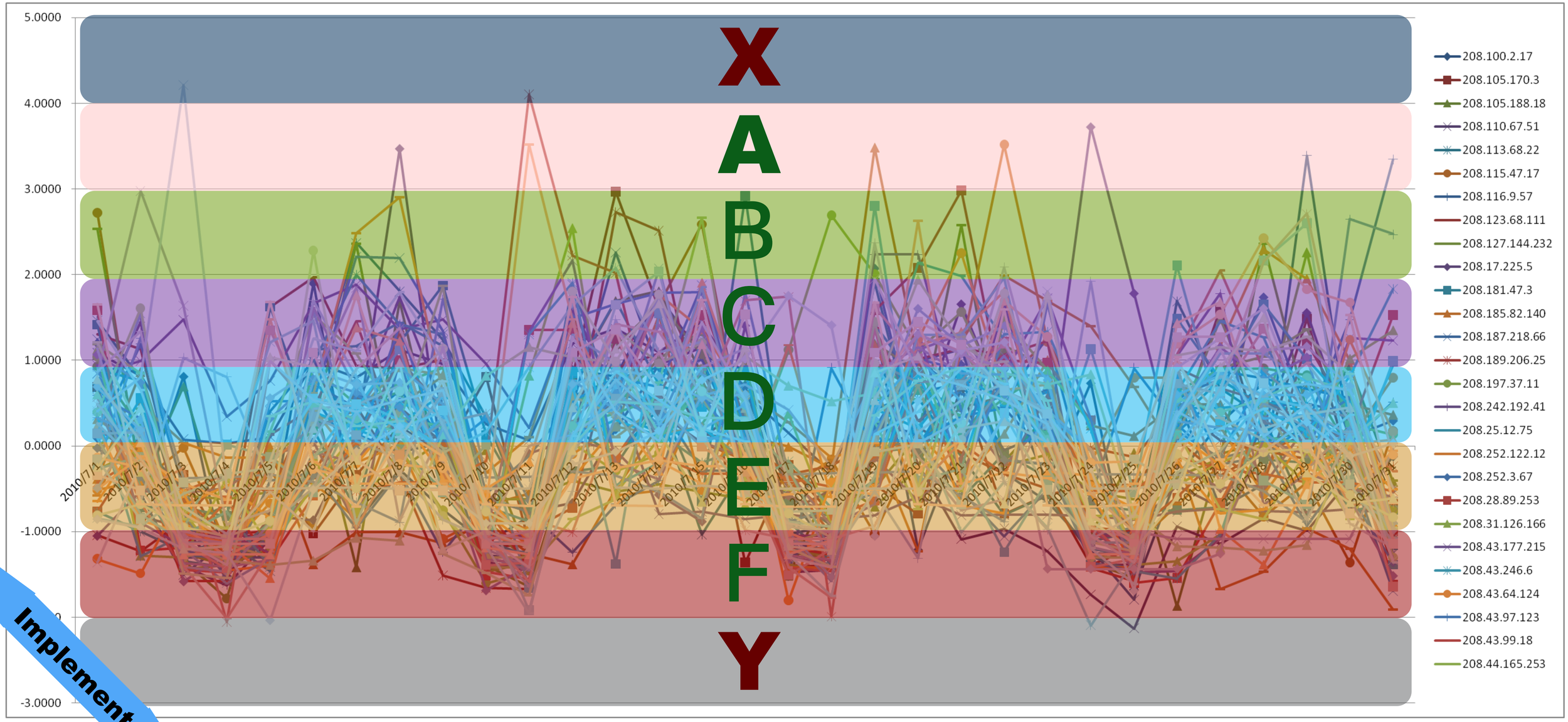
- **Problems**
  - **Similar shapes**
    - **Some shapes have a large number of queried volume; some have a little**
  - **Shifted shapes**
    - **Some similar shapes are shifted**
  - **Exactly match**
    - **If applying the hash comparison, it should be exactly matching in whole shape**

Solve Similar Shapes by Normalization



Implement

Solve Exact Match by Fuzz



Implement

208.85.53.43;

208.85.53.44;

208.85.53.45;

208.85.53.47;

208.85.53.49;

208.85.53.5;

208.85.53.43;DCDDDCCCDDDDDBCDDDDDDDDCDDCCCXD

208.85.53.44;DDDDDXDDDDDDDBCDDDDDDDDDDDDDAABD

208.85.53.45;DDDDDDCDXDDDDDDDDBCDDDDDDDDDDDX

208.85.53.47;DCDDDDBCDDDDDDDDDBBDDBCDDDCDDDD

208.85.53.49;DCDDDADDCDDDADDDDDBCDDDDDDDDDDA

208.85.53.5;DDDDDACBDDDBCBCDDDCBBDDDDCCDDDC

Implement

馬的，又來了！怎樣找

LCS(Longest Common Sequence)

http://www.csie.ntnu.edu.tw/~u91029/LongestCommonSubsequence.html

Thinking

於是，又花1-2W時間

實作LCS演算法

當興高采烈時，這個數字...真難

4,294,967,296

Thinking

於是，又花1-2D時間

...

發呆

Thinking

但很快...想到了救星

Suffix Tree

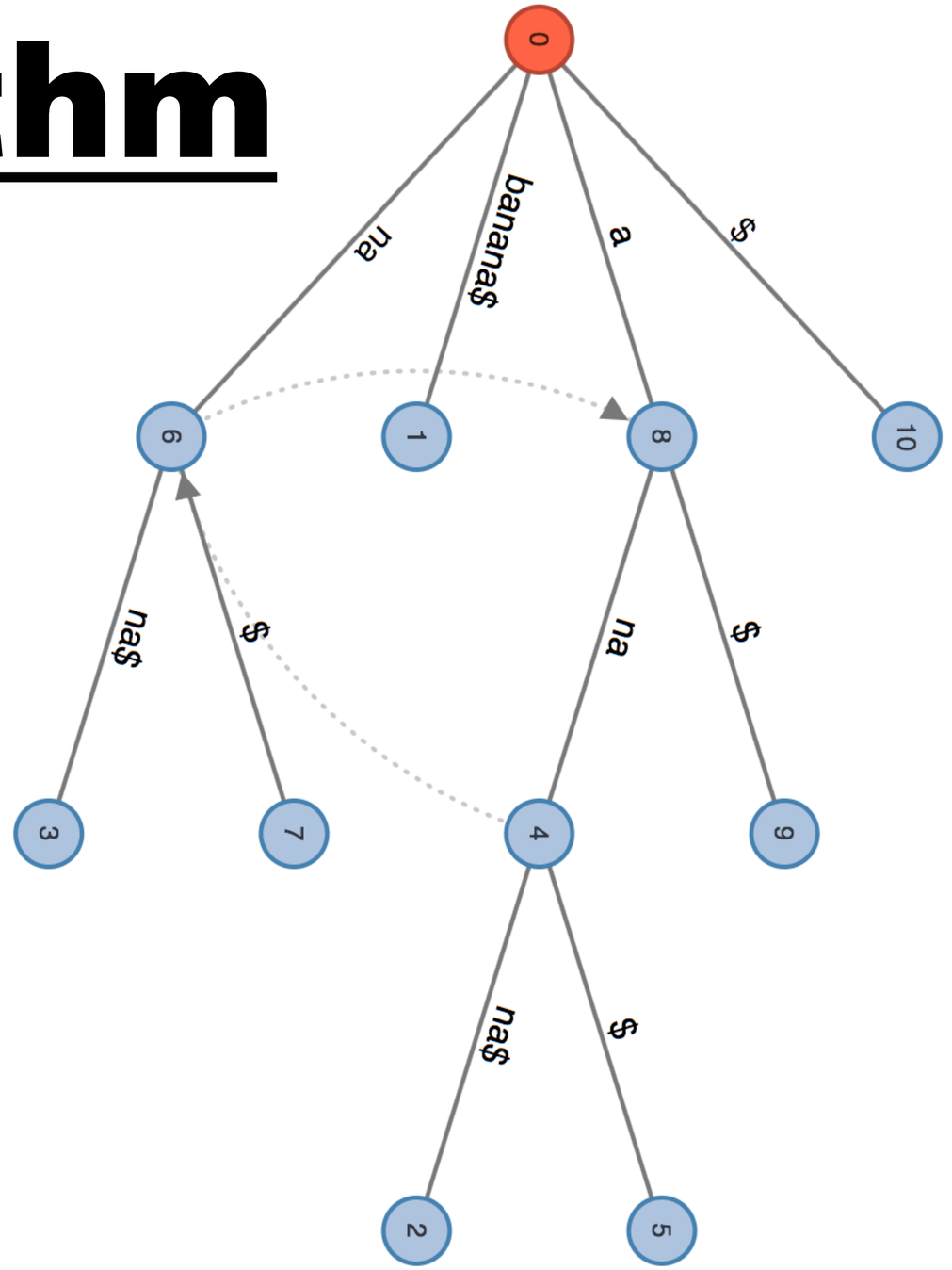
http://brenden.github.io/ukkonen-animation/

Thinking

Search Algorithm

字元種類不多但排列多

- 網頁瀏覽軌跡
  - 信用卡 > 信用卡 > 信貸 > ... > 理財
    - 有這種瀏覽軌跡的人，存款會減少的人是 x%
- 24期繳費記錄
  - 正常繳 > 正常繳 > 繳最低金額 > ... > 正常繳
    - 有這種 sequence 的人，什麼什麼比率比別人高
- 購買順序
  - 買A > 買B > 買C > 下一個最可能買什麼？



Start Cluster1: Common Sequence: **ccccceeccccceeccccceeccccce**

208.100.20.62:CCEECCCCCEECCCCCEECCCCCEECCCCE:swl;
208.11.8.10:CCEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.20.155.3:CCEFDCCCCEECCCCCEECCCCCEECCCCE:
208.228.181.118:CCEECCCCCEECCCCCEECCCCCEECCCCE:swl;
208.242.14.101:CCEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.253.82.57:CCEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.27.111.66:CCEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.27.134.5:DCEECCCCCEECCCCCEECCCCCEECCCCE:
208.27.203.66:CCEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.27.203.95:CDEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.33.52.25:CCEEECCCCEECCCCCEECCCCCEECCCCE:swl;
208.38.59.78:DDEECCCCCEECCCCCEECCCCCEECCCCE:swl;
208.38.59.82:EDEECCCCCEECCCCCEECCCCCEECCCCE:swl;
....
....

We predict that these IPs are potential GOOD IPs

Thinking

減少 71%FP

增加 US8554907

Thinking

Engineer

1. Understand your current task
2. Know your number
3. Figure out your number
4. Find out the pattern
5. Just do it

NEXT STEP

Data Scientist