

Background

Observation

Traffic Analysis - Cluster by daily volume

DS Team

Implement

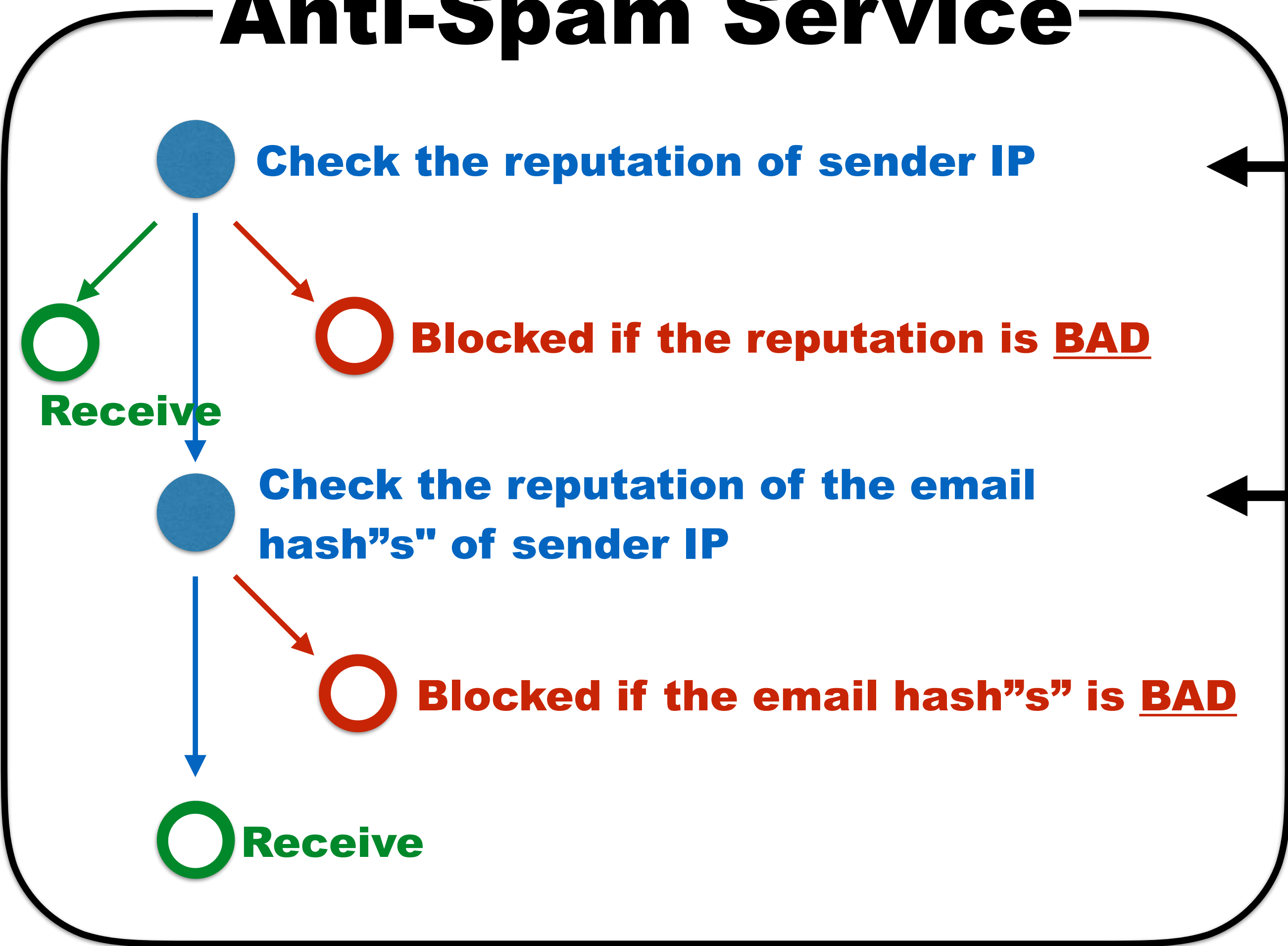
Thinking

First of All

Anti-Spam Solution



Anti-Spam Service



a. Sender IP

b. Email Hash"s"



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

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

Good

~

Spam Volume

Request Volume

Bad

~

Spam Volume

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

1

我們發生 False Positive 的郵件類型是哪一種

垃圾郵件最多是從哪個網斷劑出來，哪個國家最多

采襄什麼時候可以不用做簡報

day from customers

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

1

哪種語言我們做的最差

客戶是不是有週期性的症狀

哪個客戶的信件量越來越少
Miles 已經會用 vim 開發程式了

day from spam catcher

About 1.5 years

1. Develop programs and develop myself engineering skills
2. Know your number (Like 己菌)
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!!!

Good

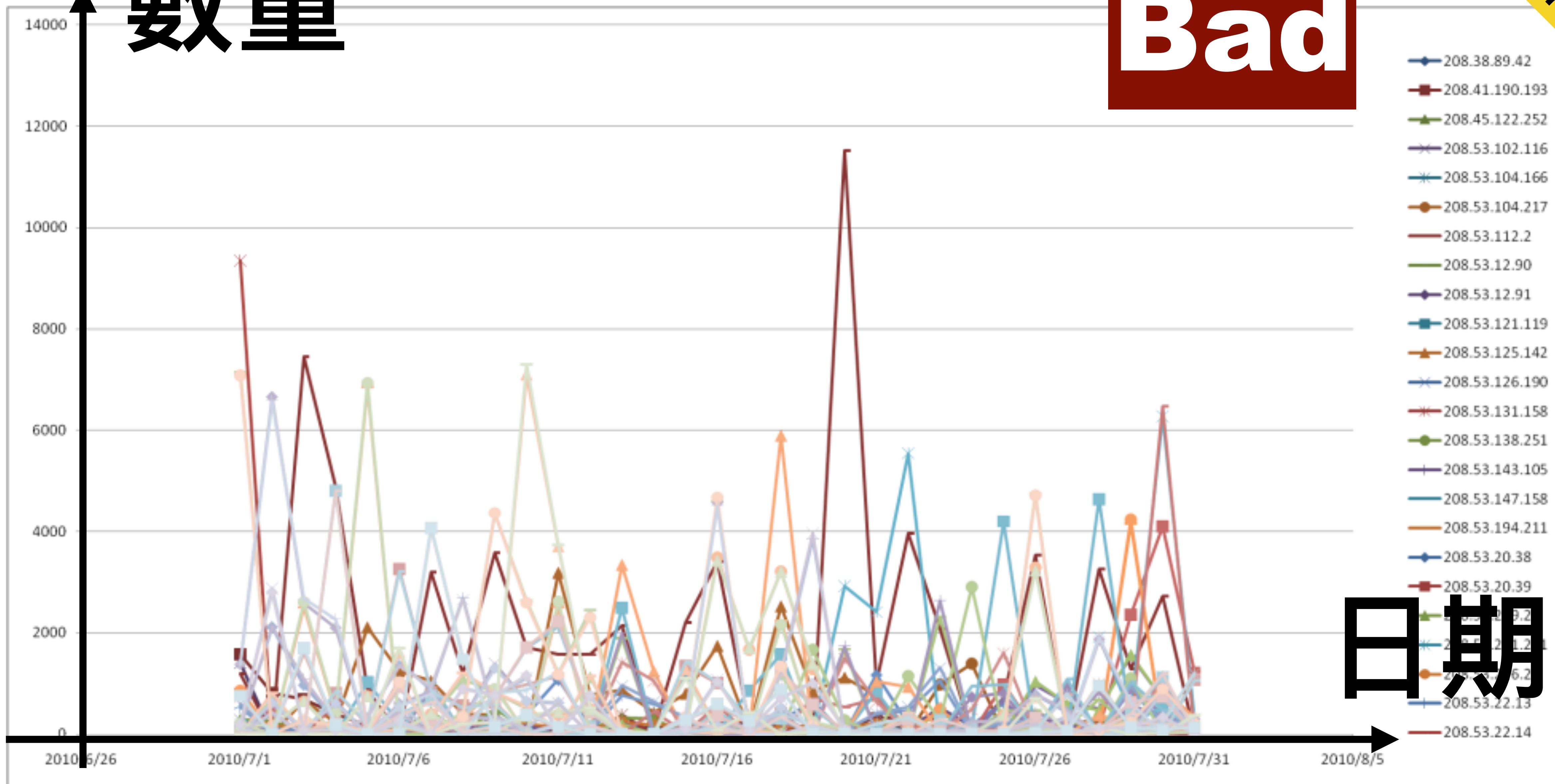


Observation

Bad

數量

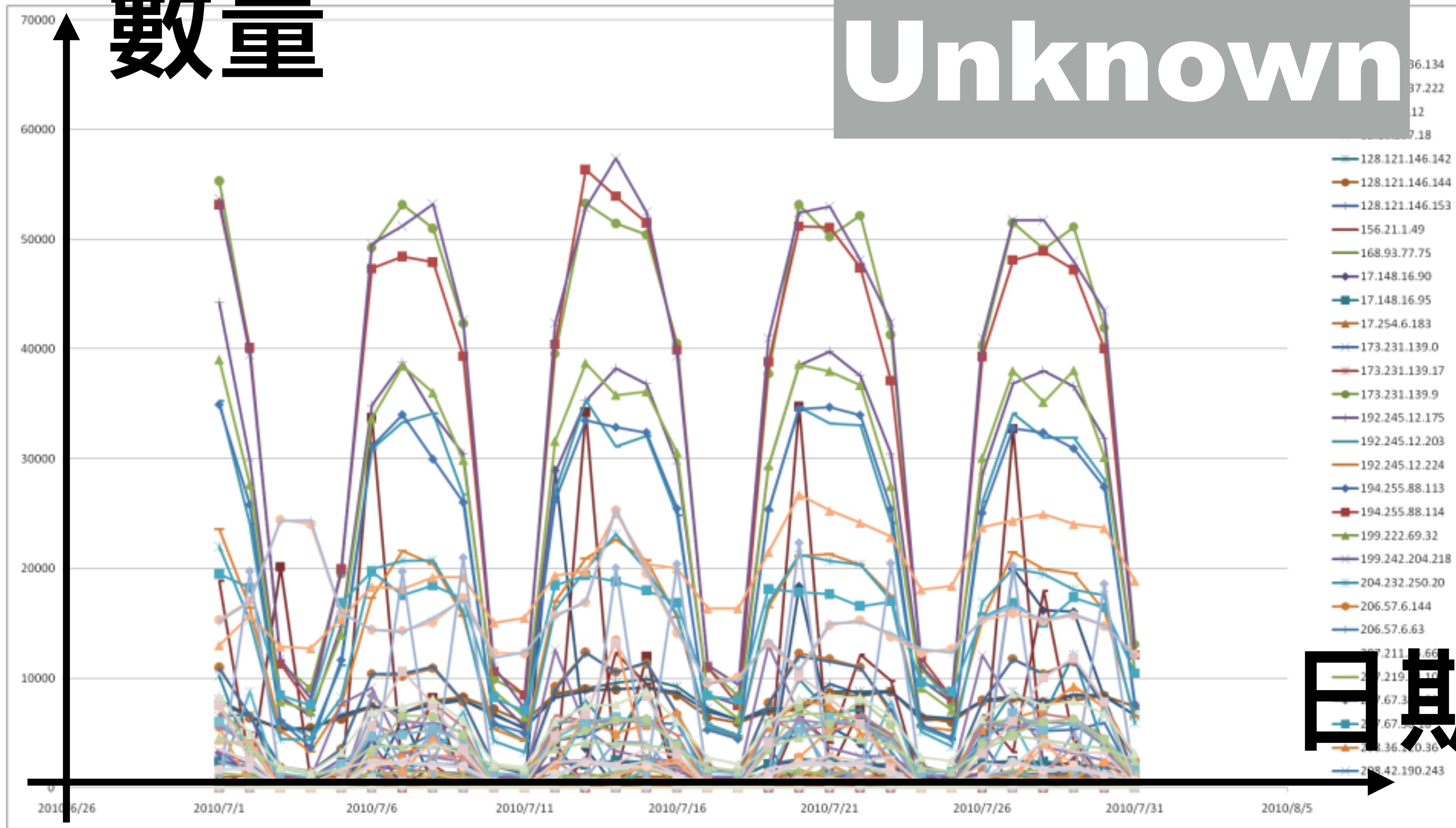
日期

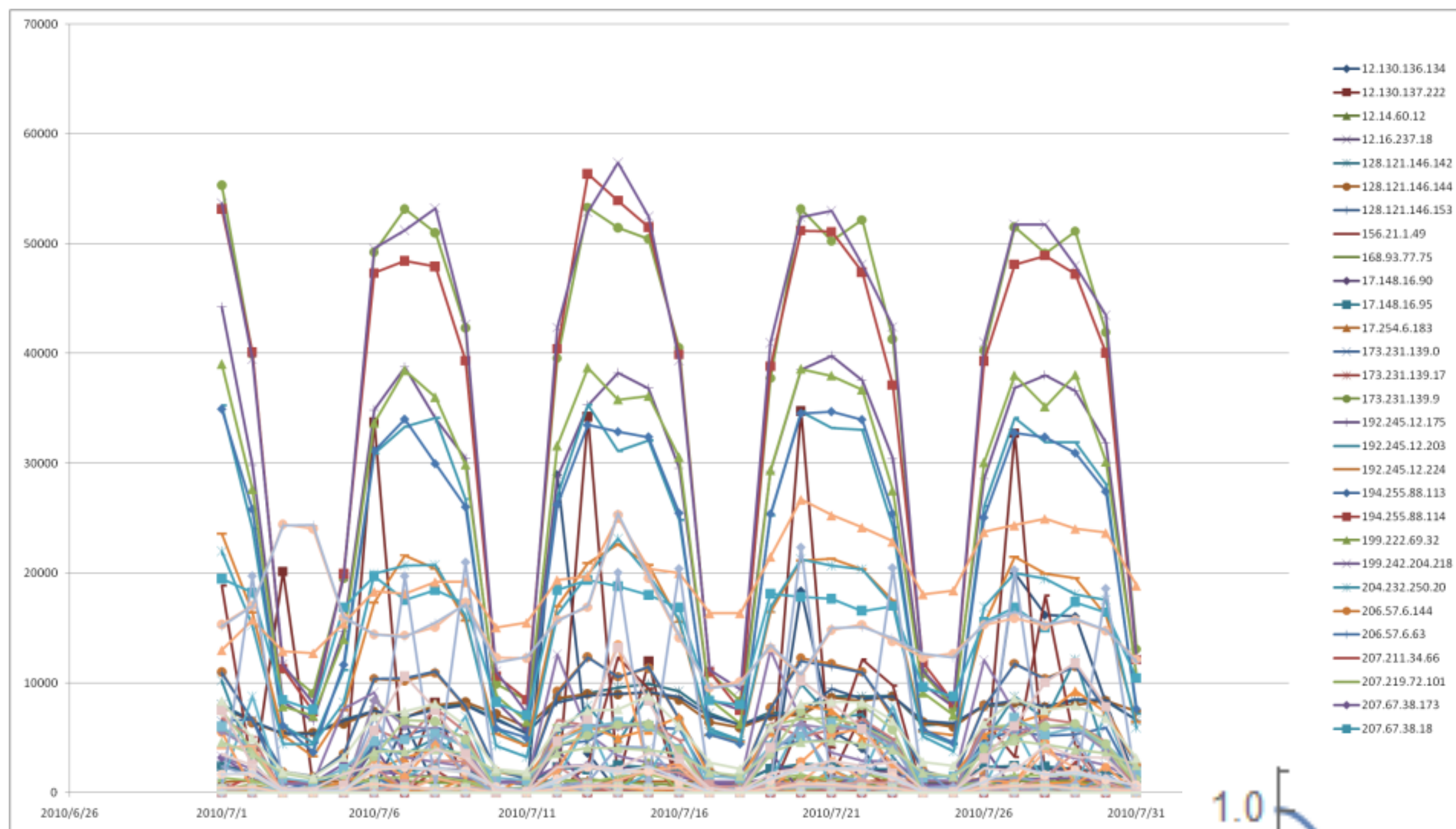


數量

Unknown

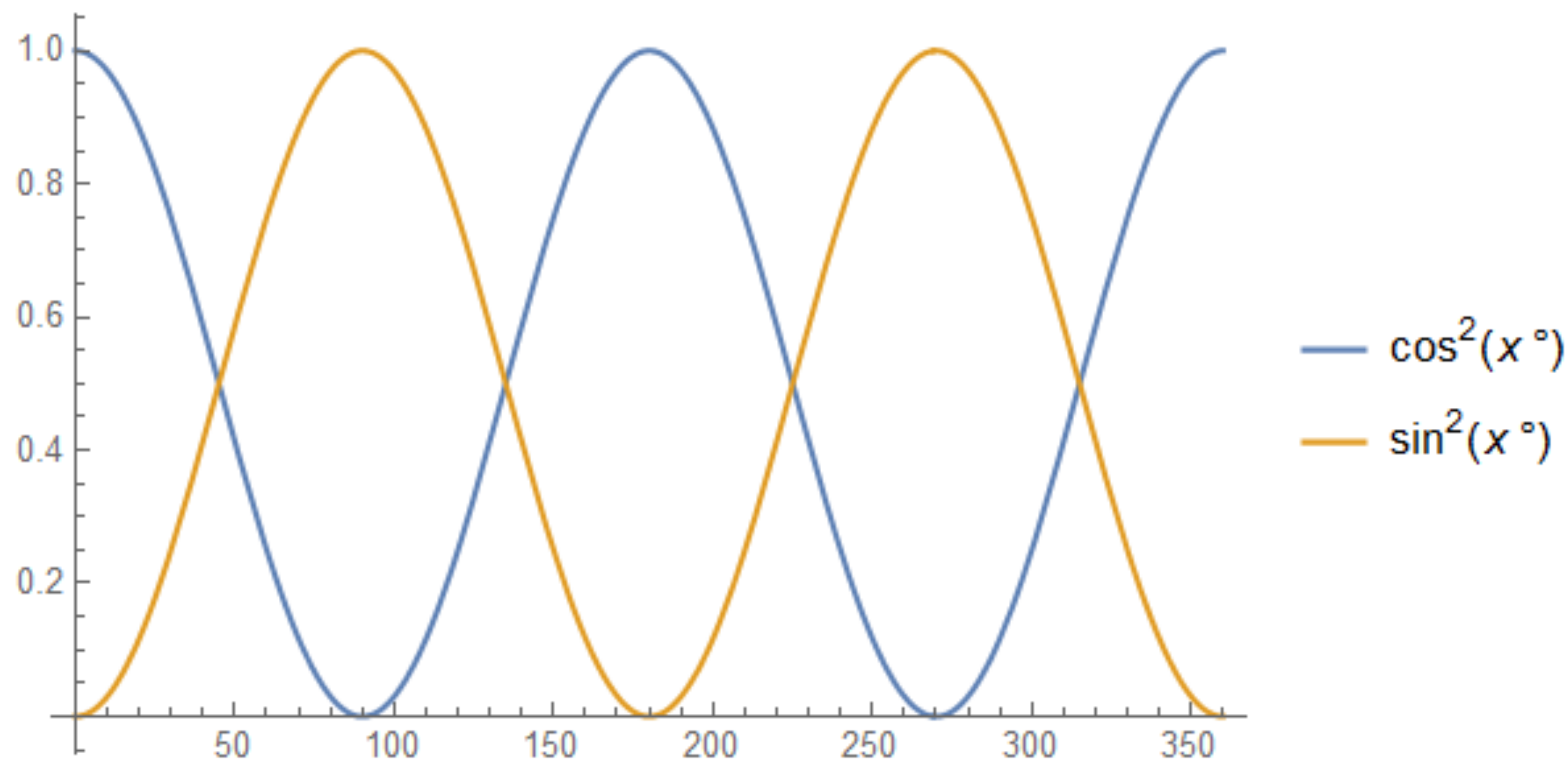
Observation





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...

但是，金庸連城訣

倒是給了靈感...

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 ...

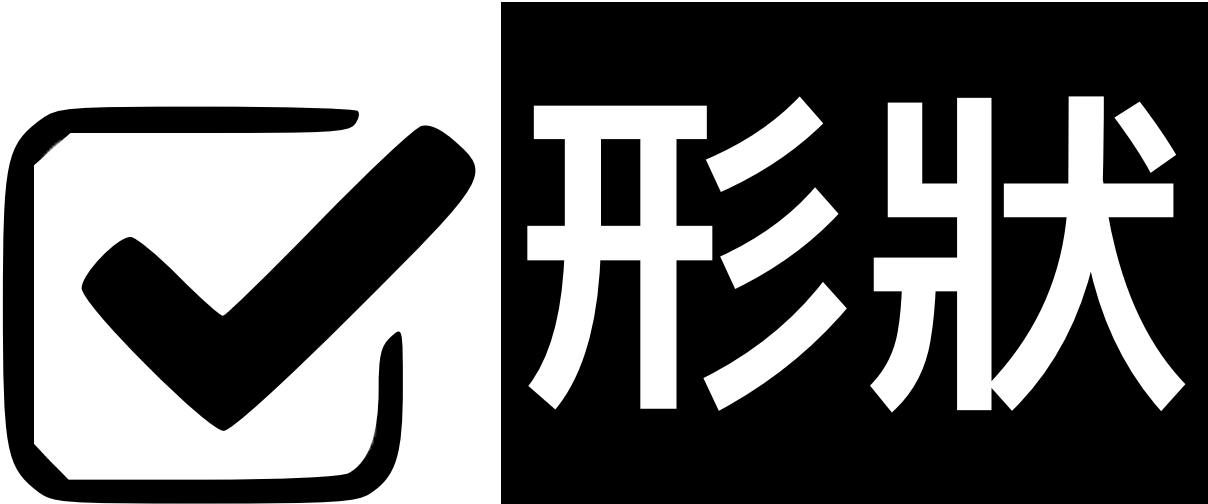
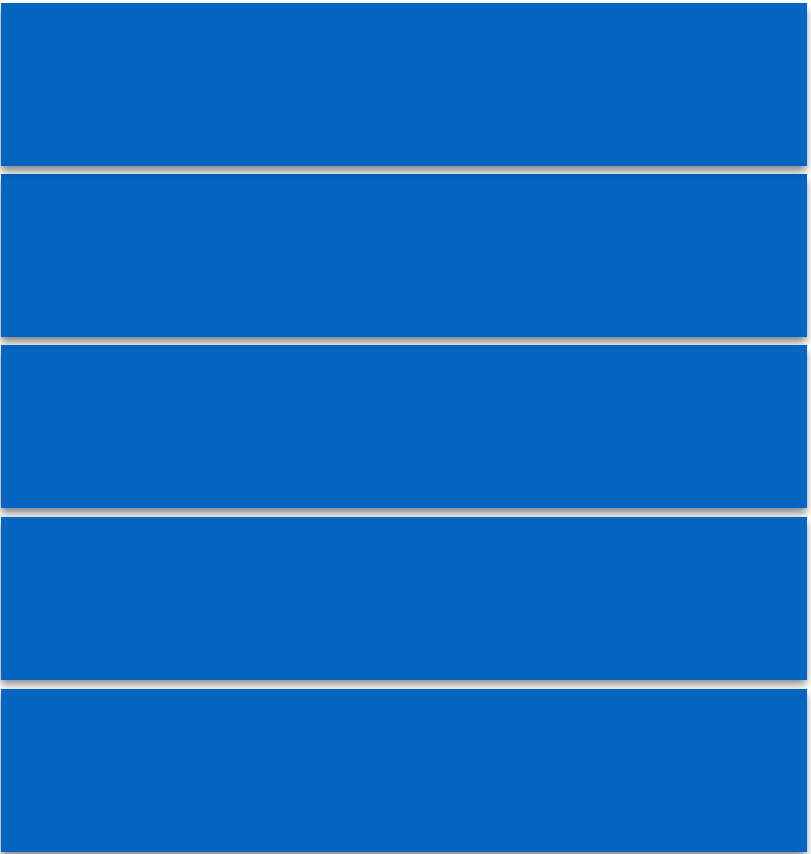
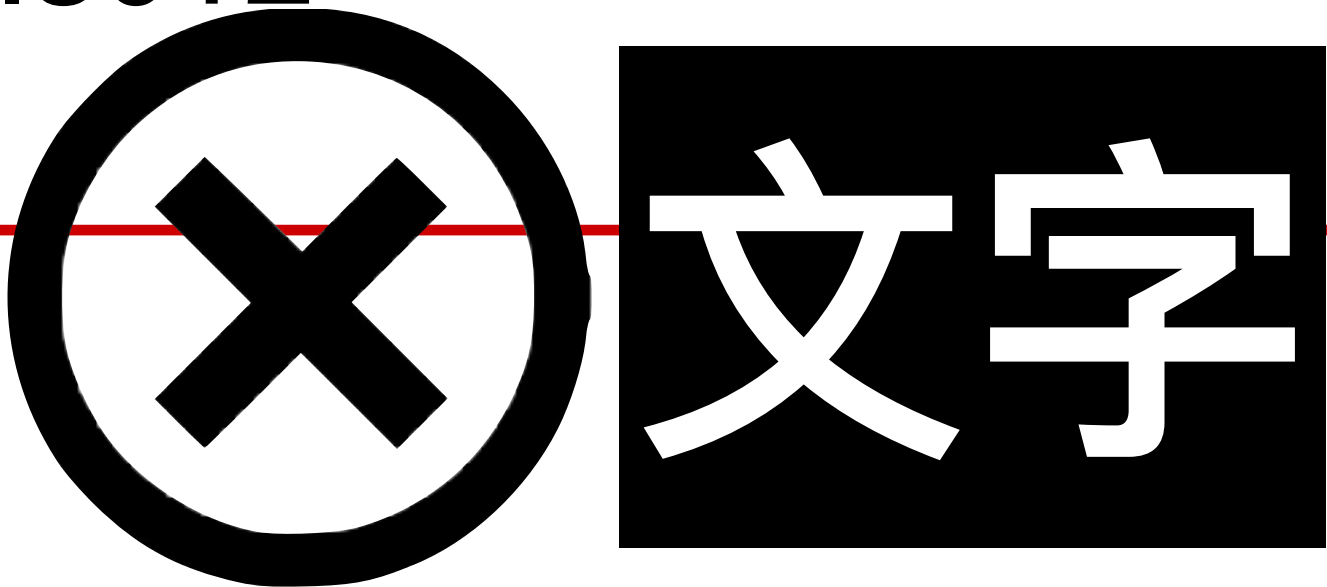
主角：石破天

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...

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
130 pills \$335.08



Implement

Cialis20mg + 4 FREEviagra Pills. No Rx

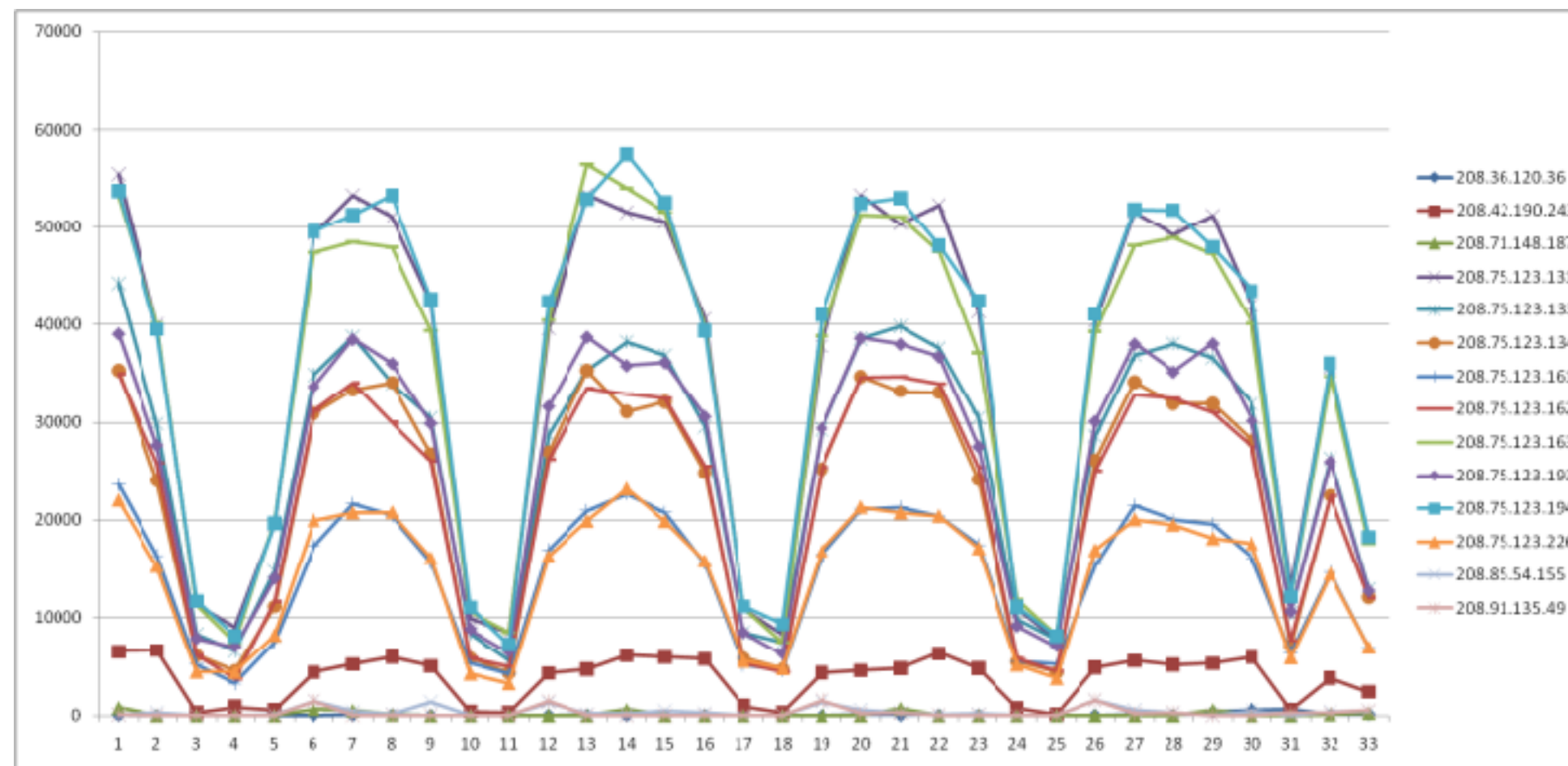
10 pills \$52.57
30 pills \$136.67
60 pills \$238.08
90 pills \$310.30
120 pills \$334.08

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

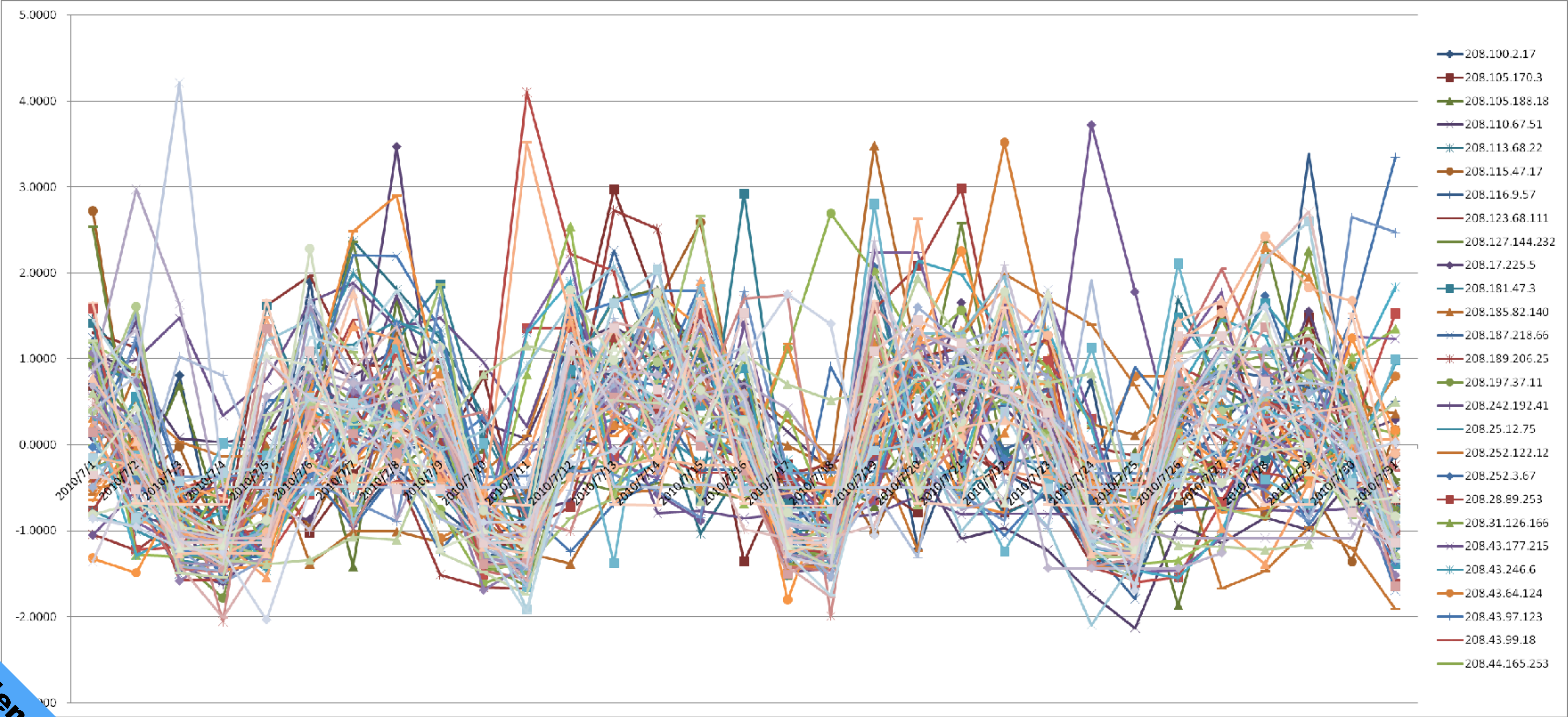


Problems

- **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

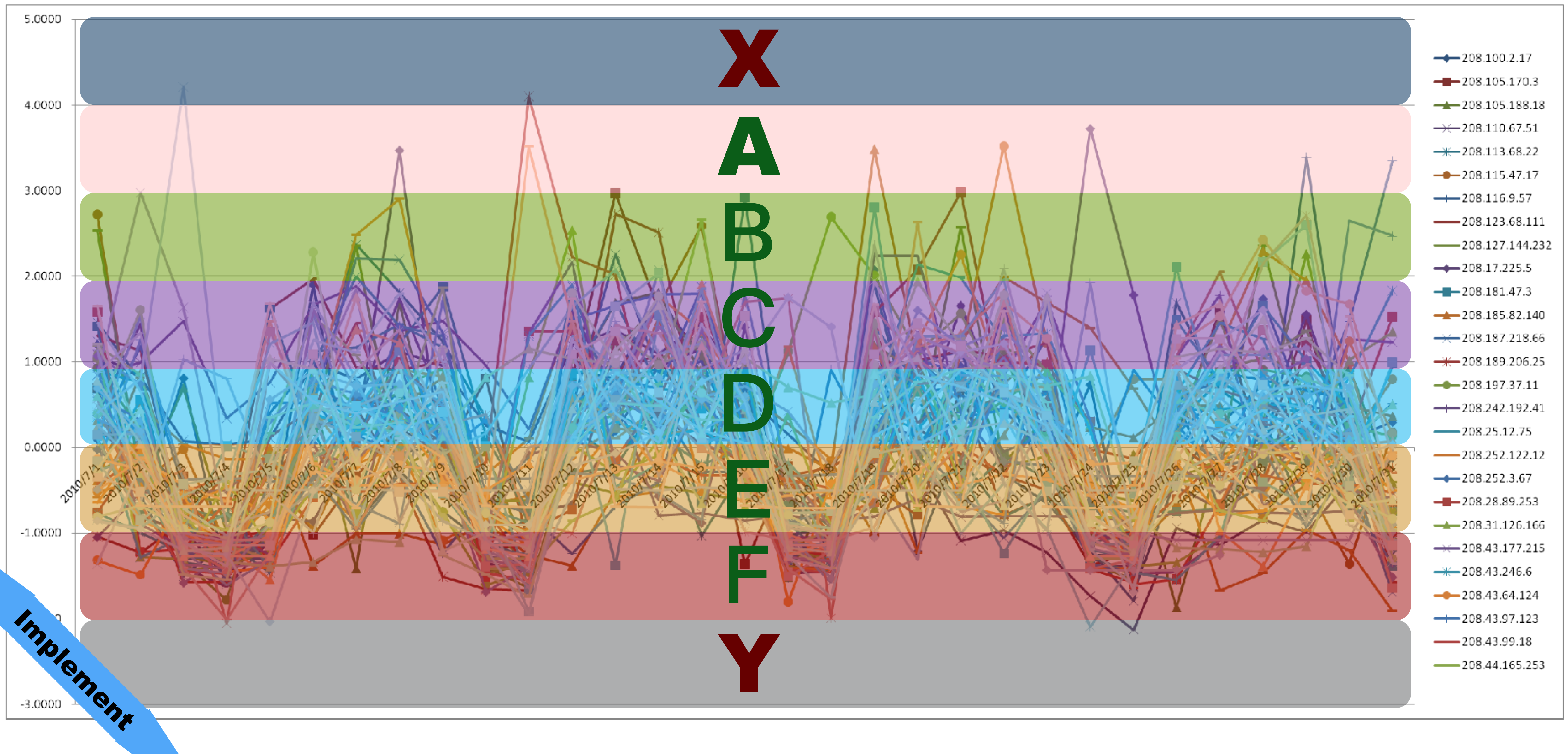
Implement

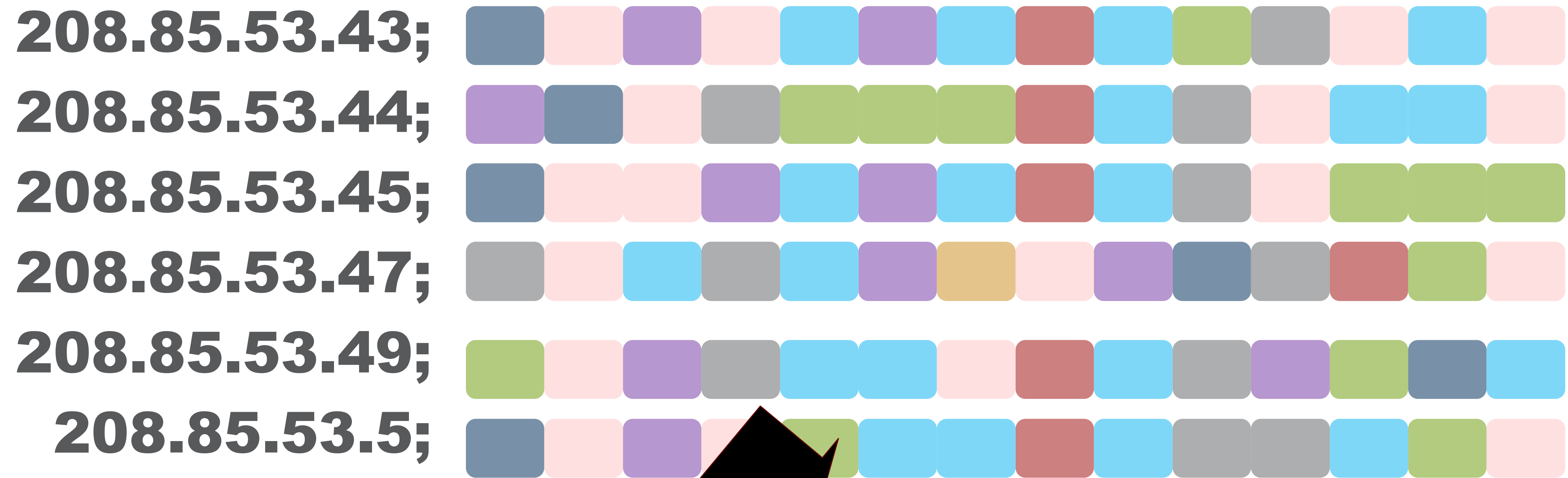
Solve **Similar Shapes** by Normalization



Implement

Solve **Exact Match** by Fuzz





208.85.53.43;DCDDDDCCCDDDDDDBCDDBDDDDDDDDDDDCDDCCCDXD
208.85.53.44;DDDDDXDDDDDDDDDBCDDBDDDDDDDDDDDDDAABD
208.85.53.45;DDDDDDDCDXDDDDDDDDDBCDDBDDDDDDDDDDDDDX
208.85.53.47;DCDDDDDBCDDBDDDDDDDDDBBDDBCDDDCDDDD
208.85.53.49;DCDDDAADDCDDDAADDDBCDDBDDDDDDDDDDDA
208.85.53.5;DDDDDACBDDDBCBBCDDDCBBDDDDCCDDDC

Implement

馬的，又來了！怎樣找 LCS(Longest Common Sequence)

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

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

於是，又花**1-2D**時間

...

發果

Thinking

但很快...想到了救星

Suffix Tree

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

Search Algorithm

字元種類不多但排列多

- 網頁瀏覽軌跡

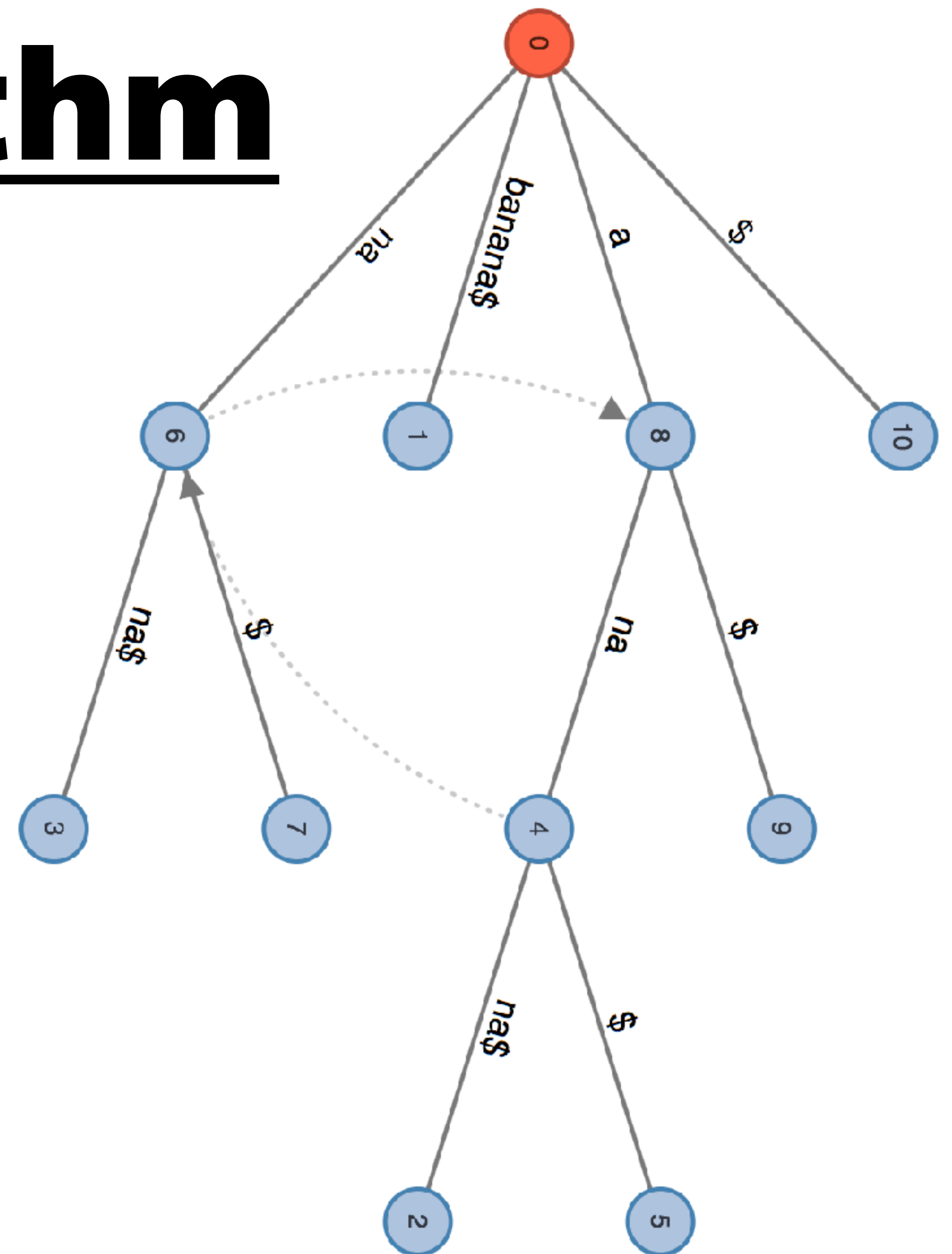
- 信用卡 > 信用卡 > 信貸 > ... > 理財
 - 有這種瀏覽軌跡的人，存款會減少的人是 **x%**

- 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: **cccceccccceccccceccccce**

208.100.20.62:CCEECCCCCECCCCCECCCCCECCCCCE:swl;
208.11.8.10:CCEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.20.155.3:**CCEFDCCCCCECCCCCECCCCCECCCCCE:**
208.228.181.118:CCEECCCCCECCCCCECCCCCECCCCCE:swl;
208.242.14.101:CCEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.253.82.57:CCEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.27.111.66:CCEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.27.134.5:**DCEECCCCCECCCCCECCCCCECCCCCE:**
208.27.203.66:CCEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.27.203.95:CDEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.33.52.25:CCEEECCCCCECCCCCECCCCCECCCCCE:swl;
208.38.59.78:DDEECCCCCECCCCCECCCCCECCCCCE:swl;
208.38.59.82:EDEECCCCCECCCCCECCCCCECCCCCE:swl;

....
....

We predict that these IPs are potential GOOD IPs

Thinking



減少

71% FFP

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



Data Scientist



**NEXT
STEP**