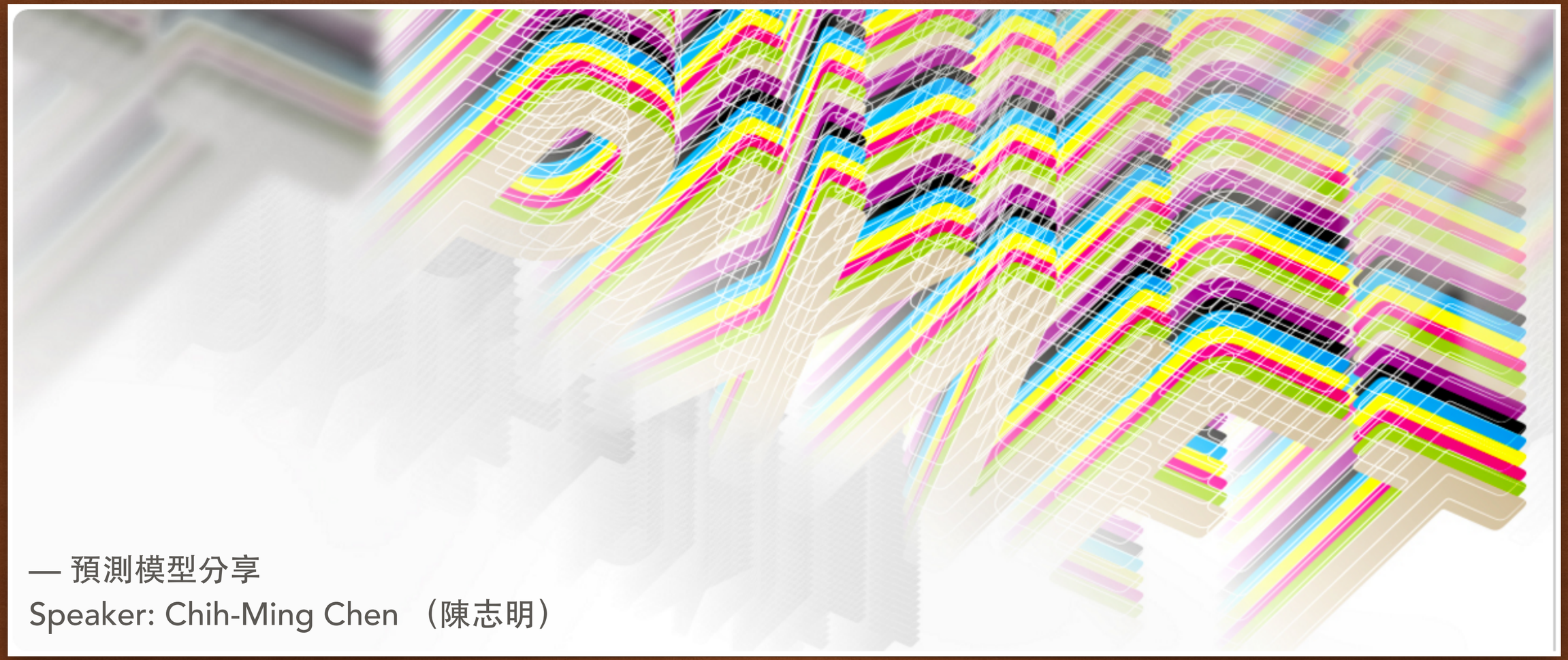
預測每位網站訪客的造訪次數 (PAGE VIEWS)

# 痞客邦 VISITOR LOG 資料挑戰賽



— 預測模型分享

Speaker: Chih-Ming Chen （陳志明）

# ABOUT ME

## CHIH-MING CHEN (陳志明)


http://clip.csie.org/~cmchen/

## Ph.D. student in TIGP-SNHCC

Advisor:
Prof. Ming-Feng Tsai （蔡銘峰）

## Research Assistant in AS

Advisor:
Dr. Yi Hsuan Yang （楊奕軒）

## Research Intern in KKBOX Inc.

Research, ML team

# 關於比賽內容

問題型態？預測目標？評估方式？

用過去的資料猜測
下一個十天的造訪次數

如何用一個最簡易的方法
取得baseline?

那就取平均吧！

結果0分...

造訪次數
Page Views

數值差距範圍大

什麼樣數值好預測？

觀察資料分布

大部分數值偏小

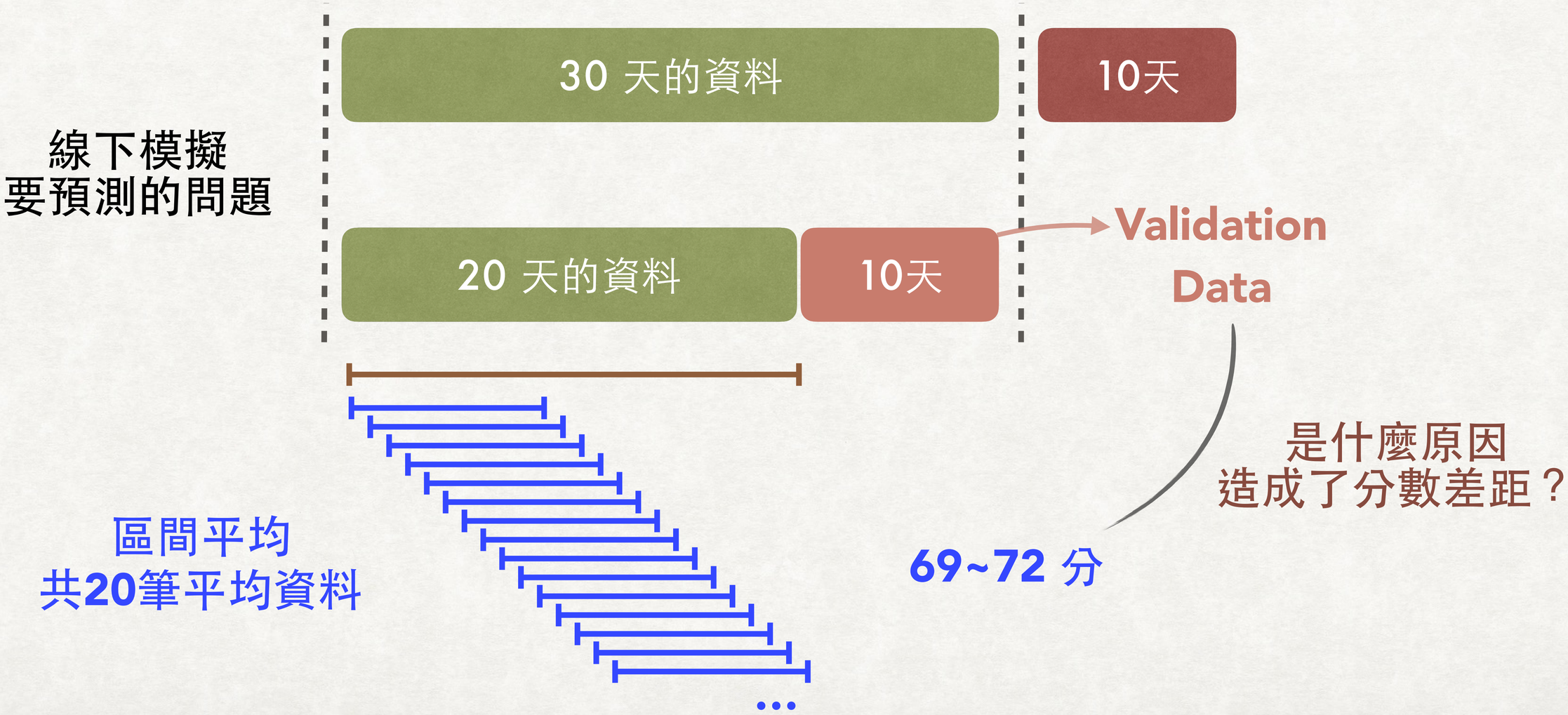
同一天集中造訪數值

KL Divergence

怎樣的策略可以達到
不錯的成績？

# 利用 Validation Data

建造一個有答案的線下測試資料



# 資料觀察

瞭解資料才能有更好的分析與策略

- url_hash - 去識別後的部落格文章 url
- resolution - 瀏覽裝置的螢幕解析度
- browser - 瀏覽裝置的瀏覽器
- os - 瀏覽裝置的作業系統
- device_marketing - 瀏覽裝置的產品型號
- device_brand - 瀏覽裝置的品牌名稱
- **cookie_pta - 去識別化的瀏覽者代碼**
- **date - 瀏覽日期**
- author_id - 文章作者 ID 去識別碼
- category_id - 文章分類
- referrer_venue - 訪客來源（網域）

**PageView = 119**

**使用者A：**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 119 | 0 | 0 | 0 | 0 | 0 | 0 |

→ **0 (outlier)**

**使用者B：**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 11 | 8 | 0 | 10 | 24 | 18 | 19 | 0 | 19 |

→ **1x**

# 偵測Outlier

## 更仔細的處理何謂 Outlier $\mathcal{N}(\mu, \sigma)$

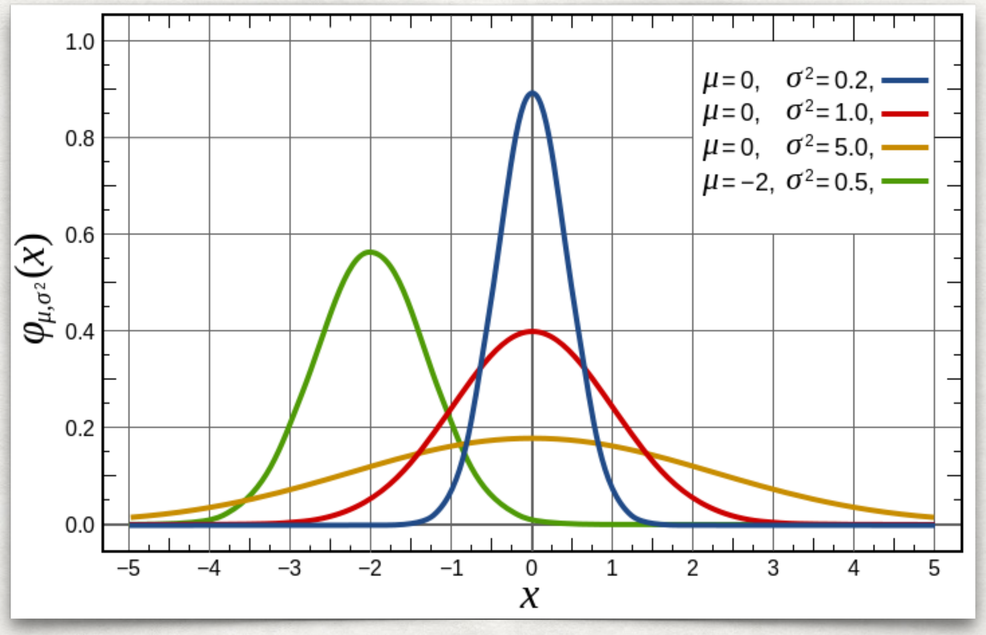
- ✓ 不同的"平均數值"使用不同倍數的"Threshold"
- ✓ 不同的"標準差數值"使用不同倍數的"Threshold"

簡單的交叉測試即可找到合適的組合 （即暴力法測試）

- ✓ 利用有拜訪天數來輔助分類

使用者拜訪網頁次數的穩定性

from wikipedia



# PSEUDO CODE

簡單就是美

```
for each user:
    compute mean, var over 30 days
    for pageview in each day:
        pageview <= min(mean+threshold*var, pageview)
    compute avg. period pageview
```

不同狀況採用不同threshold

簡單的演算法 ≠ 容易想到

# 總結

## 未來可研究方向

去除 Outlier

+

平均區間平均

可視為一個有效的特徵
給進階的模型做學習

**~80 分**

- url_hash - 去識別後的部落格文章 url
- resolution - 瀏覽裝置的螢幕解析度
- browser - 瀏覽裝置的瀏覽器
- os - 瀏覽裝置的作業系統
- device_marketing - 瀏覽裝置的產品型號
- device_brand - 瀏覽裝置的品牌名稱
- **cookie_pta - 去識別化的瀏覽者代碼**
- **date - 瀏覽日期**
- author_id - 文章作者 ID 去識別碼
- category_id - 文章分類
- referrer_venue - 訪客來源（網域）

許多尚未開發的資料屬性

還有很多進步空間！！

可使用什麼樣的
機器學習模型？

可代入什麼
模型預測的技巧？

可做什麼
有價值的延伸應用？

# ANY QUESTION?

*Chih-Ming Chen (陳志明)*
*changecandy at gmail.com*