那兩年，我們一起追的 Hadoop

Outline

- 資料總覽
- 資料前處理
  - Missing Value
  - Clean Data
- 預測模型建立
- 系統架構與參數
- 結語

資料總覽

	View	Search	Cart	Order	Total
Train	4,696,645 (86%)	337,306 (6%)	368,450 (7%)	58,922 (1%)	5,461,323 (100%)
Test	4,662,896 (93%)	374,964 (7%)	-	-	5,037,860 (100%)

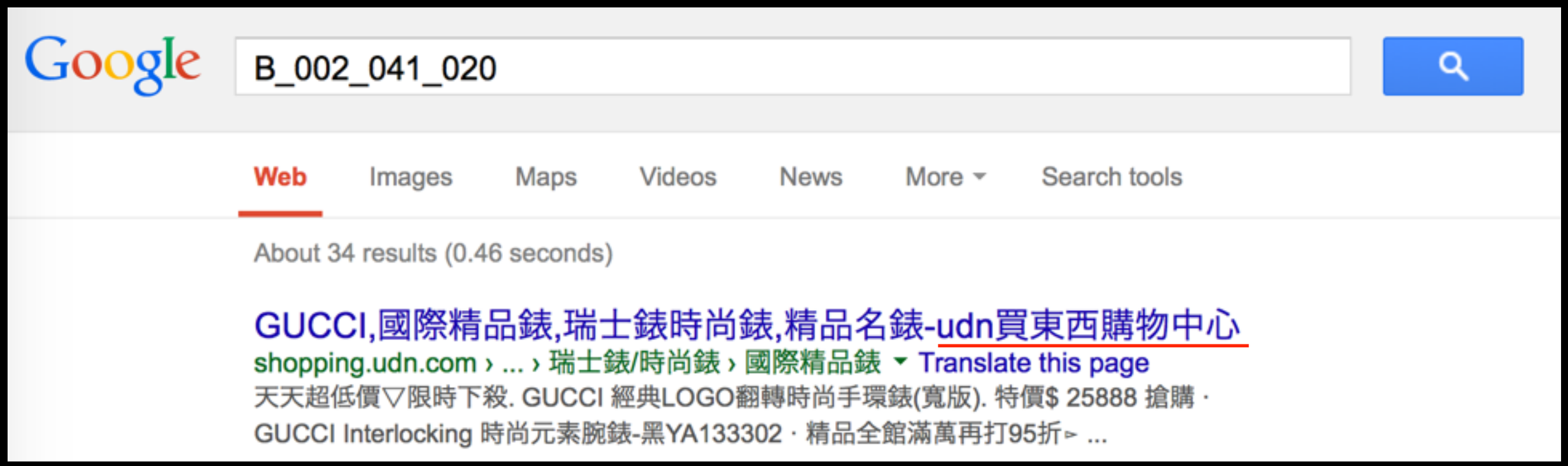
Know data first

Where the dataset comes from?

Category list

```
P/1.1" 302 160 "-" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) C
hrome/29.0.1547.62 Safari/537.36"
114.36.11.187 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=view;uid=U129297265;pid=
0023468384;cat=B,B_002,B_002_041,B_002_041_020;erUid=e88c3e9a-7e55-76a9-3f89-70c6b334cba;
 HTTP/1.1" 302 160 "-" "Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.2; WOW64; Triden
t/6.0; Touch; MAARJS)"
1.164.129.191 - - [01/Feb/2015:00:00:01 +0800] "GET /action?;act=view;uid=;pid=0009827053
;cat=E,E_002,E_002_021,E_002_021_007;erUid=1d4c20b8-d54c-46ef-1330-3de831a920b2; HTTP/1.1
" 302 160 "-" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome
:
```

Just google it



Missing Values

Product Missing Values statistics

Dataset	Without price	Without
Train	188,059 (77%)	21,038 (9%)
Test	203,508 (87%)	0 (0%)
Train + Test	251,732 (82%)	21,038 (7%)

Total product items are 308,682

Product ids convention

- Product id format in udn is different with dataset, we need to convert them into udn format.

Digits	Product Id	udn product id	Comment
10	1234567890	U123456789	
13	1234567890ABC	U123456789	product variant

Interpolation for product price based on pid range and category

Before Interpolation

```
000001015,188,A_004_017_003,台灣製
000001017,188,A_004_017_003,台灣製
000001018,188,A_004_017_003,台灣製
000001022,188,A_004_017_003,台灣製
000001023,198,A_004_017_003,台灣製
000001033,188,A_004_017_003,台灣製
000001034,188,A_004_017_003,台灣製
000001070,0,A_004_017_001,
000001072,220,A_002_017_015,創意工
000001097,1790,F_017_011_002,iNen
000001105,1490,F_017_008_005,iNen
000001108,1730,F_017_008_005,iNen
000001110,1299,F_017_008_003,iNen
```

After Interpolation

```
000001015,188,A_004_017_003,台灣製
000001017,188,A_004_017_003,台灣製
000001018,188,A_004_017_003,台灣製
000001022,188,A_004_017_003,台灣製
000001023,198,A_004_017_003,台灣製
000001033,188,A_004_017_003,台灣製
000001034,188,A_004_017_003,台灣製
000001070 220,A_004_017_001,
000001072,220,A_002_017_015,創意工
000001097,1790,F_017_011_002,iNen
000001105,1490,F_017_008_005,iNen
000001108,1730,F_017_008_005,iNen
000001110,1299,F_017_008_003,iNen
```

Interpolation for product category based on pid range and price

Before Interpolation

```
000001282,599,G_023_001_012,聯統-10吋手
000001284,14800,
000001285,11800,G_006_002_011_002,安體
000001286,14800,G_006_002_011_002,安體
000001287,14800,G_006_002_011_002,安體
000001290,14800,G_006_002_011_002,安體
000001292,8800,G_006_002_011_002,安體百
000001293,8800,G_006_002_011_002,安體百
000001294,8800,G_006_002_011_002,安體百
000001295,8800,G_006_002_011_002,安體百
```

After Interpolation

```
000001282,599,G_023_001_012,聯統-10吋手
000001284,14800,G_006_002_011_002,
000001285,11800,G_006_002_011_002,安體
000001286,14800,G_006_002_011_002,安體
000001287,14800,G_006_002_011_002,安體
000001290,14800,G_006_002_011_002,安體
000001292,8800,G_006_002_011_002,安體百
000001293,8800,G_006_002_011_002,安體百
000001294,8800,G_006_002_011_002,安體百
000001295,8800,G_006_002_011_002,安體百
```

Product Missing Values statistics

Dataset	Without price	Without
Train + Test (original)	251,732 (82%)	21,038 (7%)
Train + Test (pid convention)	240,594 (78%)	3,444 (1%)
Train + Test (crawling udn)	3,827 (0.1%)	501 (0.2%)
Train + Test (Interpolation)	0 (0%)	0 (0%)

Total product items are 308,682

Data Cleaning

Category Extraction

- Category log format is inconsistency

Category type	Log	Log after extraction
multiple cids	cat=J,J_007,J_007_009, J_007_009_016	cat=J,J_007,J_007_009, J_007_009_016
one cids	cat=H_004_017_004	cat=H,H_004,H_004_017, H_004_017_004

Date Shift

Dataset	Date	Date after shift
Train	2015/2	2013/9
Test	2015/3	2013/10

- Train DS has iPhone 5, iPad 4 and iPad mini without iPhone 5s
- Test DS appears iPhone 5s when date is 2013/10/29~2013/10/31

iPhone 5s、5c台灣售價公布25日上市- 手機動態聚焦 ...
mag.udn.com/mag/vote2007-08/storypage.jsp?f... - Translate this page
Oct 23, 2013 - udn數位資訊：聯合新聞網經營之資訊頻道，整合原「數位玩樂誌」、「數位文化誌」 ... iPhone 5s、iPhone 5c預計將在10月25日於台灣市場正式銷售，目前台灣 ... 兩款新機均預計在10月25日在台上市，但台灣官網並未如第一波上市國家 ...

Decode Search Keyword

```
36.230.39.90,2015-02-01 00:10:51,,３ｍ　隱形,ff0ff75f-e8ac-9ddb-bf1b-aa449822b085
1.34.131.167,2015-02-01 00:10:57,,gucci|包,dc945994-2472-2cf5-2fdc-eb85defc5465
220.137.3.34,2015-02-01 00:10:58,U234579365,mp3,92e720da-17be-2b67-3383-9e5ccbd9499f
218.166.6.240,2015-02-01 00:11:05,,落健,3227e323-71df-70fd-efec-dc03e856ad07
118.161.204.147,2015-02-01 00:11:07,U398804258,電腦桌,8e253a92-1e43-480-b644-79441bf03b8a
61.58.168.22,2015-02-01 00:11:19,,奧利佛,189c0c8-6509-776d-d3f7-d4fb7208b200
1.162.43.249,2015-02-01 00:11:20,,創見,66ed4b7-bfd4-8432-2d85-ee0880326f07
1.34.131.167,2015-02-01 00:11:26,,gucci|夾,dc945994-2472-2cf5-2fdc-eb85defc5465
1.169.33.96,2015-02-01 00:11:33,,6632,eab7f756-4ace-a67e-b59c-44aeefb72c16
175.180.94.248,2015-02-01 00:11:40,,好吃滷味,853c463d-4996-31e8-1fe2-f43c9b52a9bb
36.225.164.46,2015-02-01 00:11:50,,包鐘包,84689b9b-afaa-2b47-1886-4b4fbbdd91d5
36.236.115.77,2015-02-01 00:11:50,,espon|141,57fc3a0d-11dd-912b-2be6-5deb3aadebd
112.104.97.194,2015-02-01 00:11:51,,空白,e18fc7c7-490-b470-1301-9d552daf8b
123.204.126.78,2015-02-01 00:11:55,,彩虹馬,d33c7bc8-182c-dd0d-d256-89d05d97550f
36.225.164.46,2015-02-01 00:11:58,,包中包,84689b9b-afaa-2b47-1886-4b4fbbdd91d5
1.162.43.249,2015-02-01 00:12:01,,創見隨身碟,66ed4b7-bfd4-8432-2d85-ee0880326f07
218.166.6.240,2015-02-01 00:12:01,,首爾,3227e323-71df-70fd-efec-dc03e856ad07
61.58.168.22,2015-02-01 00:12:04,,奧利佛,189c0c8-6509-776d-d3f7-d4fb7208b200
106.1.53.220,2015-02-01 00:12:08,,耳麥,fb4180e4-d875-f34d-fbdb-6773ee72a199
49.158.208.247,2015-02-01 00:12:11,,Hello|Kitty,47cac58b-3311-31a-b4f-26b6af67248d
114.32.66.251,2015-02-01 00:12:11,,分裝,2fd022c8-480c-ff30-5b74-39823f4ee41f
125.230.67.137,2015-02-01 00:12:12,U46488849,eyah,d8d2d9f6-52d8-cbb5-2aef-baafd1472995
106.1.53.220,2015-02-01 00:12:15,,耳麥,fb4180e4-d875-f34d-fbdb-6773ee72a199
1.162.43.249,2015-02-01 00:12:15,,創見隨身碟,66ed4b7-bfd4-8432-2d85-ee0880326f07
114.46.139.167,2015-02-01 00:12:17,,Sandia|POLO,d07514e5-595b-d987-cd79-e2d935a2111
219.85.255.122,2015-02-01 00:12:27,U440888317,U002282725,bb96ea56-41a-75-7b32-75e88f8737
123.193.123.32,2015-02-01 00:12:32,U448908360,夏慕尼,9c48c11e-1a93-73fd-9a47-991ec83a94fc
36.233.145.46,2015-02-01 00:12:33,,傳真機,f9eb1fb0-6fca-989b-ea0f-bacaa31d6b6
```

Prediction

Problem#1

預測哪些 eruid 是有購買可能

設計思路

- 商品統計數據能對總體有個概念，但無法預測過去數據不顯著的商品
- 對使用者行為歸類，能套用在不同商品
  - view log by session -> 買 or 不買
  - view log by pid -> 買得多 or 買得少

Input
access log 依據 eruid 合併

Features	Description
erUid	使用者id
viewCount	使用者在Session看過幾次商品
uniqeViewCount	使用者在Session看過幾次不重複商品
cat[01ABCDEFGHIJKLOV]	Session中看過某類別的次數
maxCat	Session中看過多次的類別
buy	Session中是否有發生購買

Output
eruid是否有購買

erUid	buy
A	Yes
B	No
C	No

Algorithm
Random Forest

- 最初是想用 view log 的使用者行為，利用決策樹簡單的分類「買」或「沒買」
- 而 Random Forest 可以給我們較單一決策樹更好的建議
- ntree: 30

Problem#2

預測產品的購買率

Input
access log 依據 pid 合併 & erUid購買預測結果

Features	Description
pid	商品id
viewCount	商品被看過的次數
viewBySession	商品被幾個Session瀏覽過的次數
price	商品價格
cat	商品類別
buyRate	商品購買率

Output
pid的購買總數

pid	buyRate
A	6.1
B	1.7
C	8.8

Algorithm
SVM

- SVM 是相當適合做數據預測的工具，使用先前介紹的欄位去預測出 buyCount。
- buyCount 的預估值非整數，實際運用我們視它會 buyWeight 後續利用它來套用基因演化。
- Kernel: Radial

Problem#3

估算購買數

設計思路

- 已有先前的二個 model：判斷某個 session 會不會買、判斷某個 pid 被買多少，對於最終的結果還差商品數量
- 針對所有商品，需找出推算實際購買數量
  - {pidN} * {countN} * {price}
  - 其中 pidN 與 price 是已知的，採用 Genetic Algorithms 推算 countN

Algorithm
Genetic Algorithms

- initial Population：以 Order Model (svm) 獲得的 {pid, buyWeight} 集合為主，並加已知在榜內的 {Intop20Pid, 1.0} 作為第一代資料集
- fitness function：
  - 已知答案離 top 20 越近分數越高
  - 整體參數 (representation) 分佈與 {pid, buyWeight} 越一致分數越高

tuning Genetic Algorithm

- 運用 GA 產生的最終 buyCount 去產生 top 100 清單
- 利用新的 top 100 剔除測試過的 pid 後，用 web 驗證工具找出新的解，配合商品統計數據比較，並再次加入 GA 初始參數產生新的清單

系統架構設計

效能參數說明

創意加分

Summary

- Hadoop ecosystem
  - MR - preprocessing
  - Hive - import csv dataset
  - Impala - analysis and ad-hoc query
- Machine Learning
  - Weka - model validation
  - R - Random forest and svm
  - Java - Genetic algorithm
- CLI
  - q - ad query (early development)

Conclusions

- 拋棄成見
  - 常拿著統計數據與預測結果在猶豫「真的有人買這東西嗎？」
  - Machine Learning 的黑盒子，推薦了個上個月 800 名以外的品項，主觀上不敢放入名單，但它卻是在名單之內。

Thanks :)