從電影推薦預測開始學機器學習

Mia Chang

twMVC

About me



Data Scientist

- Algorithm Research
- Machine Learning / Deep Learning

Community

- twMVC
- Core Staff R-Ladies
- Co-Founder Azure Taiwan Community
- Co-Founder Tech Podcast Night

Contribute

- 2017 Data Platform MVP
- MVP Award Blog Technical Committee
- Workshop / Hackathon Mentor

About me

簡單介紹推薦系統

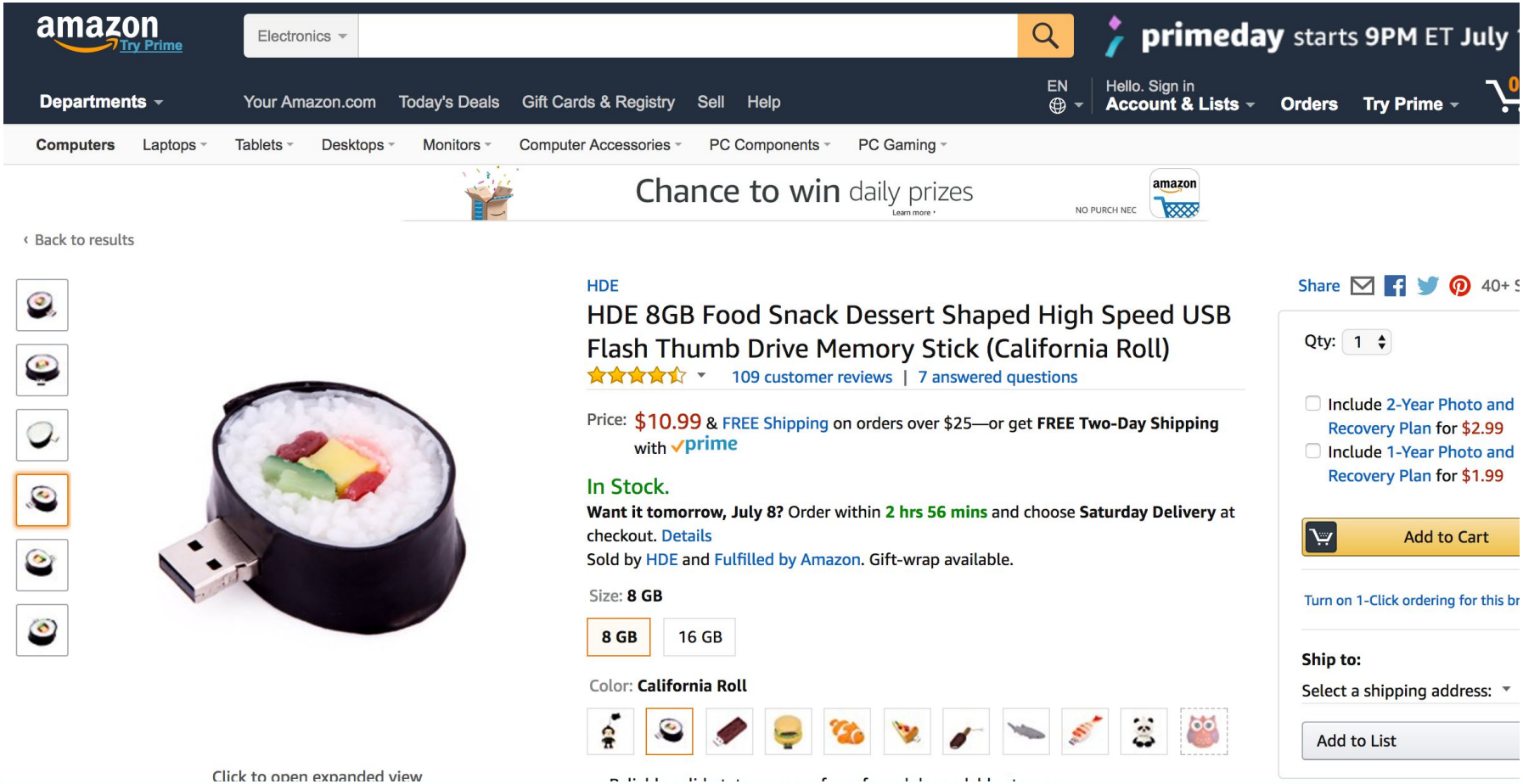
操作示範 [實驗、API]

Recap, 複習



簡單介紹推薦系統

關於推薦系統 Amazon

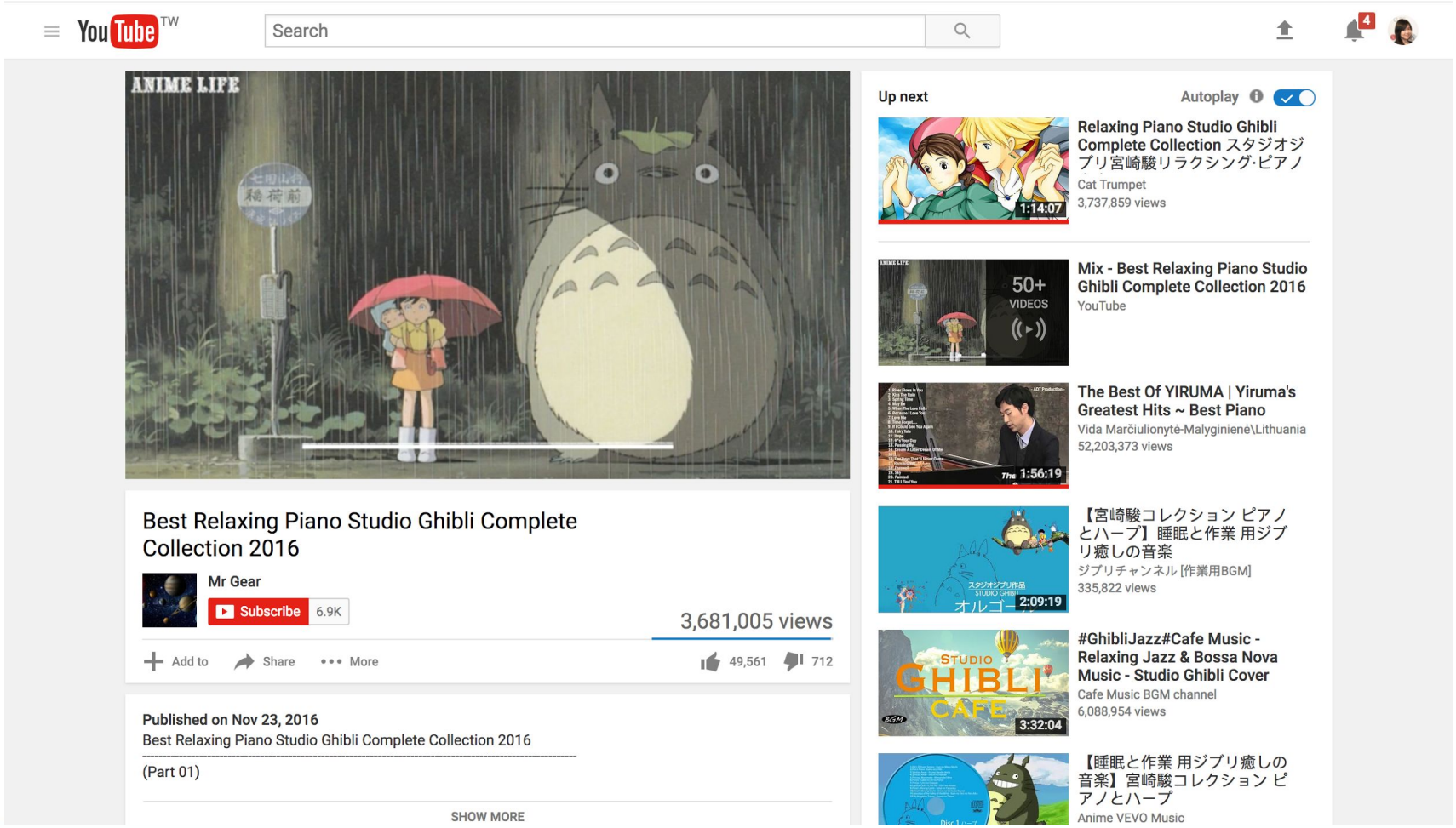


關於推薦系統 PChome



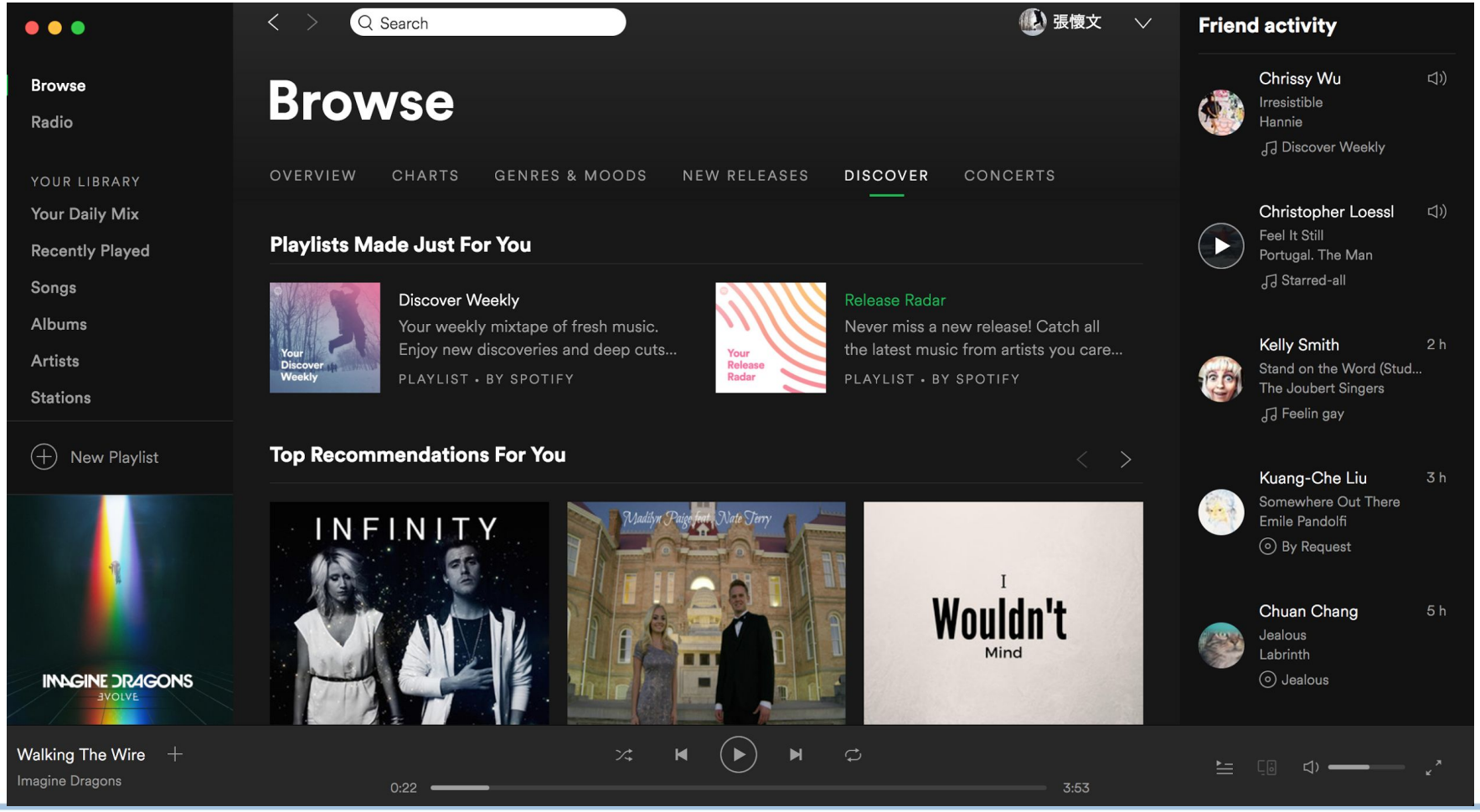
ref: PChome 截圖

關於推薦系統 YouTube



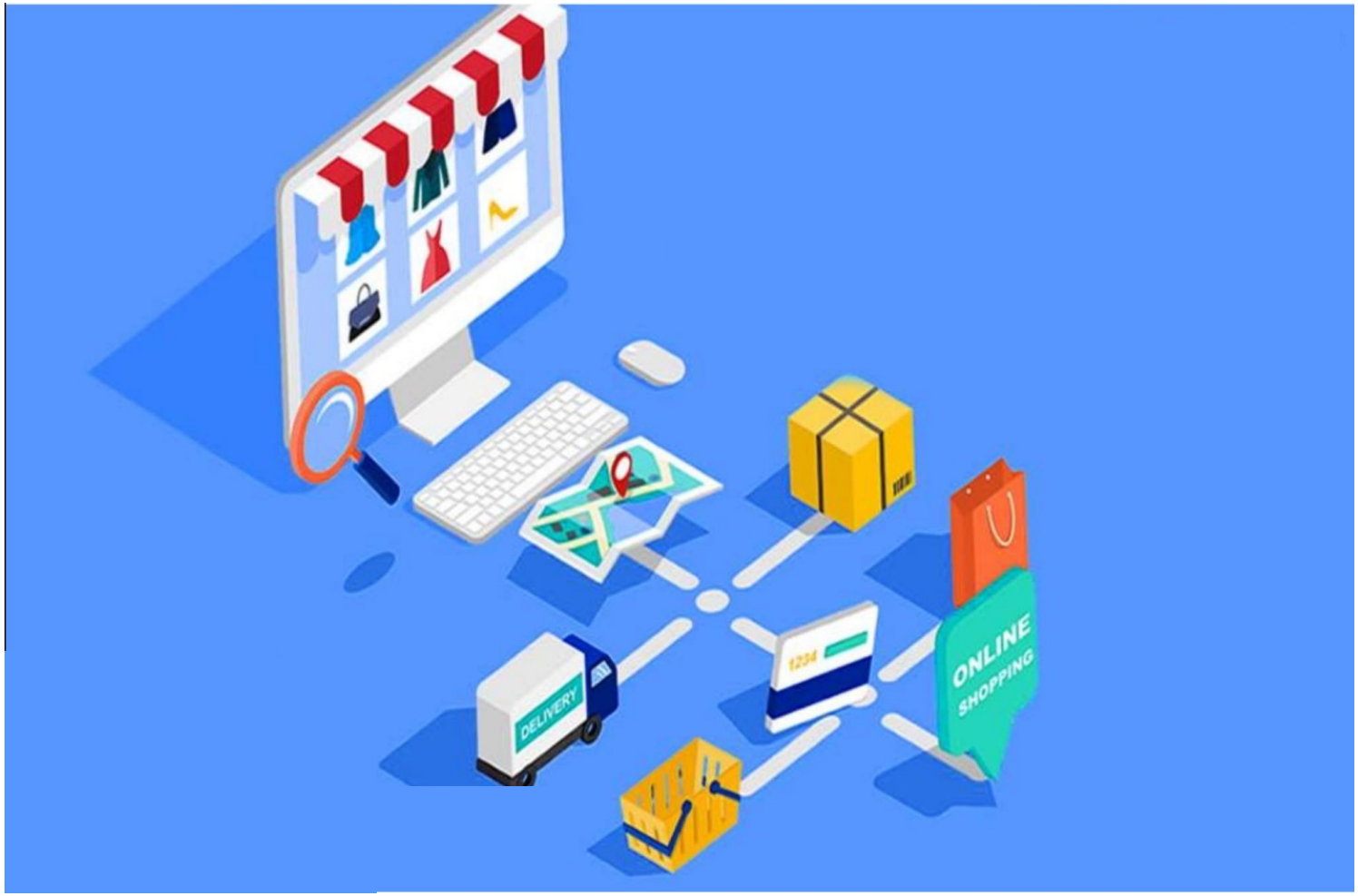
ref: YouTube 截圖

關於推薦系統 Spotify



ref: Spotify 截圖

什麼時候適合建立推薦系統



ref: https://www.slideshare.net/shopieasy

什麼時候適合建立推薦系統

- 站在商家的立場
  - 宣傳新品，增加曝光
  - 透過推薦拉長顧客停留時間
- 站在顧客的立場
  - 在太多商品的情境下找到相似的物件
  - 在太多商品的情境下找到適合我的物件

推薦系統的起源:尿布與啤酒?

資料分析領域流傳這麼一則故事:

某家超市(大多時候是 Walmart)使用某種分析方式(大多時候是關聯分析)分析資料庫資料,發現每週五晚上啤酒與尿布銷量有正向關聯性。

原因是星期五晚上年輕父親會幫家裡買尿布,並順便為週末球賽買啤酒。

ref: 啤酒尿布是都市傳說

推薦系統的起源:尿布與啤酒?



ref: 啤酒尿布是都市傳說



定義分析問題

問題介紹

- 我們想要解決_____的問題。
  - 透過手上的_____資料。
  - 我們想要產生一個_____的模型。

問題介紹

- 我們想要解決____商品無法被相對應使用者看見____問題。
  - 透過手上的____使用者行為____資料。
  - 我們想要產生一個____推薦系統____的模型。

有了要解決的問題

接下來思考...

什麼是machine learning?

一個問題是如何被描述的？

什麼是machine learning?





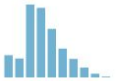
ref: https://www.youtube.com/watch?v=cKxRvEZd3Mw

Feature

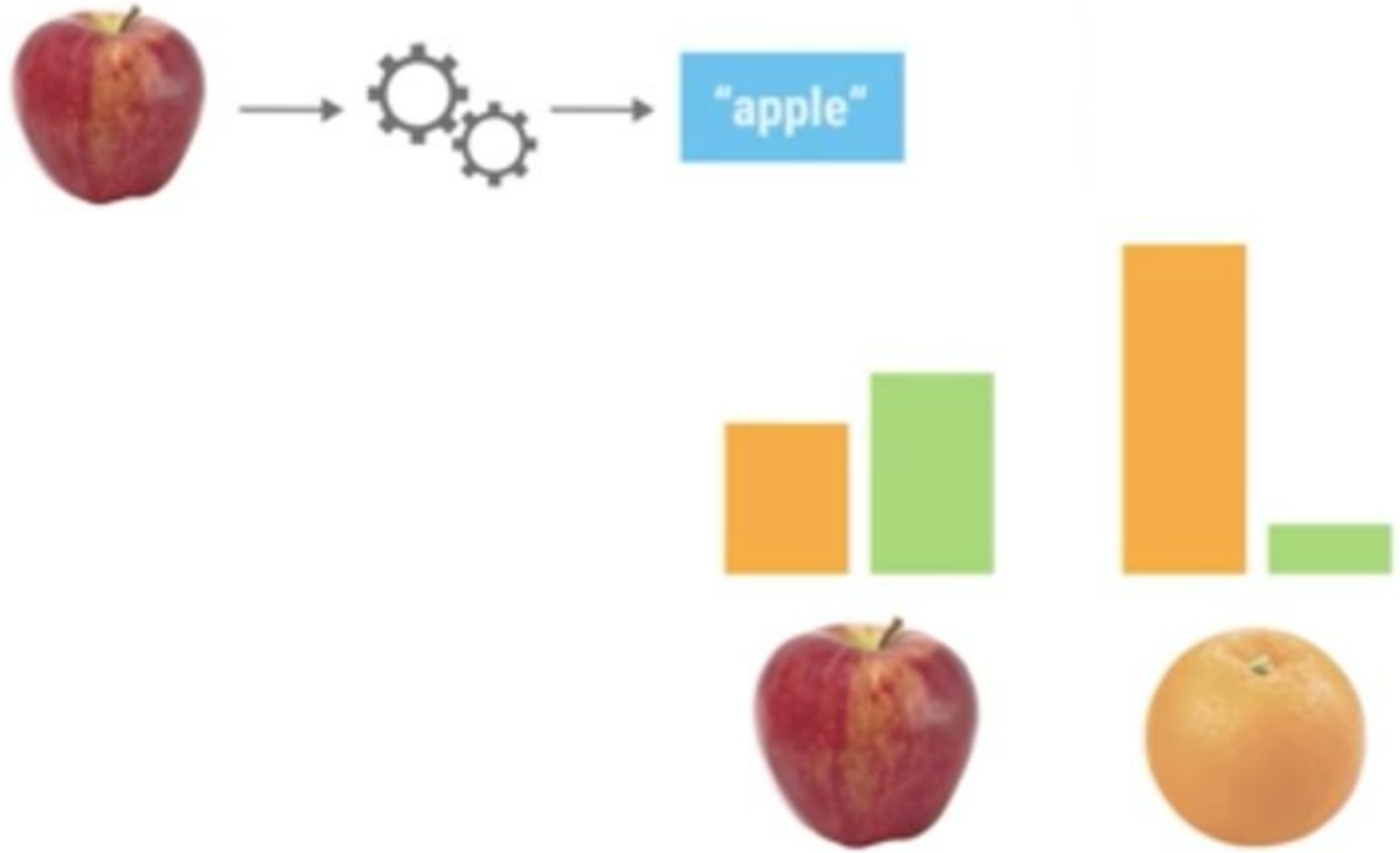
rows 891

columns 12

view as

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
								
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 211
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 175
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/ 310128
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803

什麼是machine learning?



資料集



- Analysis from MovieTweetings

  225,000 ratings for

  15,742 movies

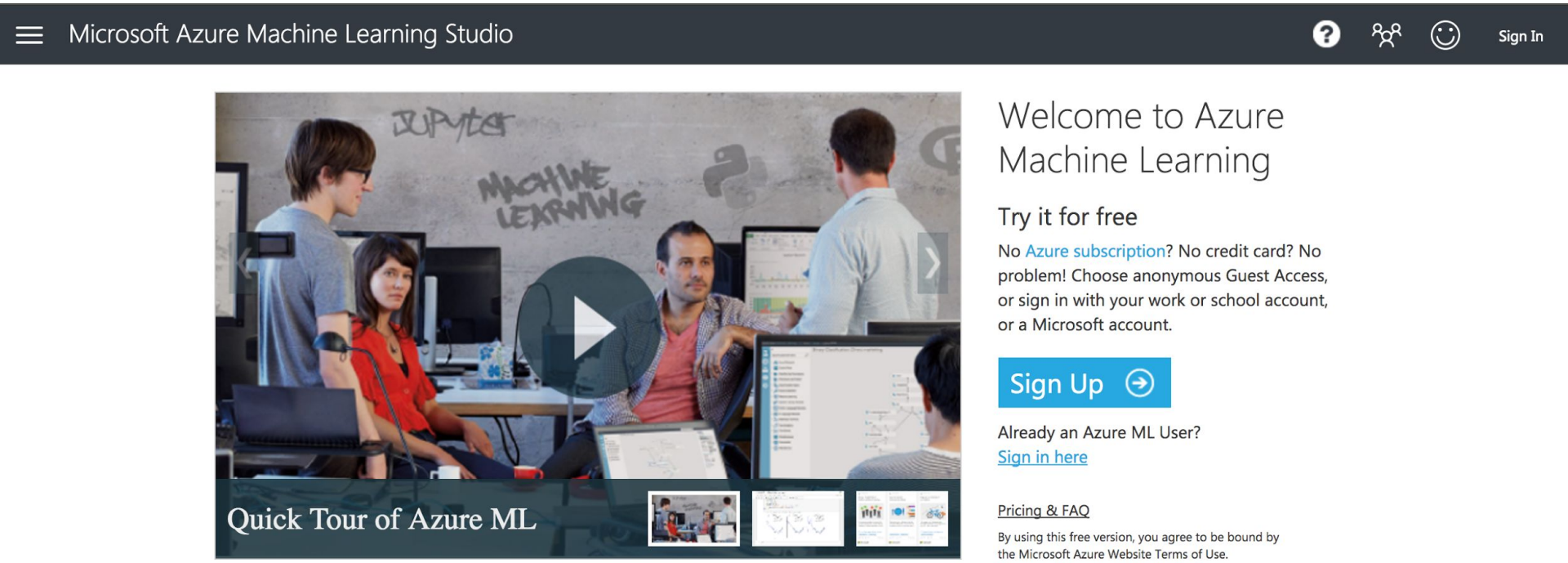
  by 26,770 users



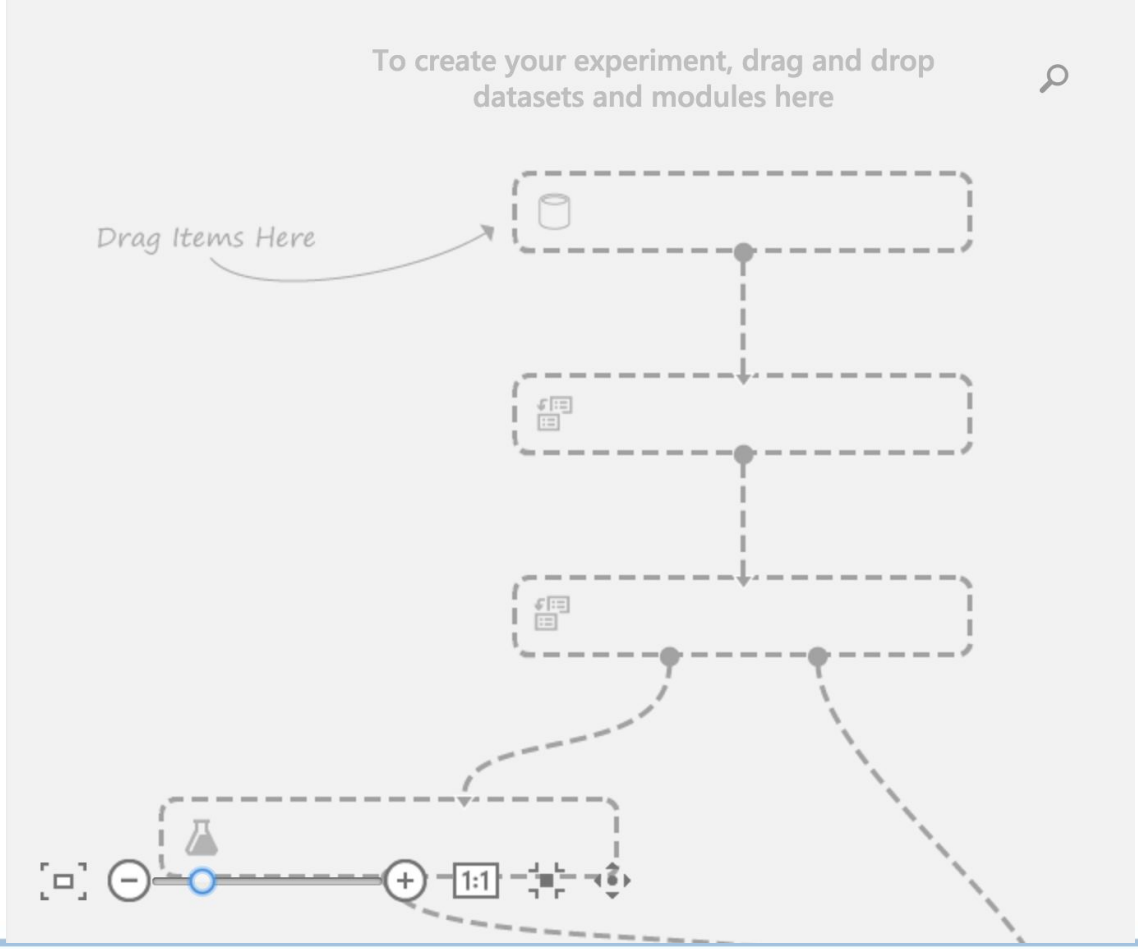
Demo

Azure Machine Learning Studio

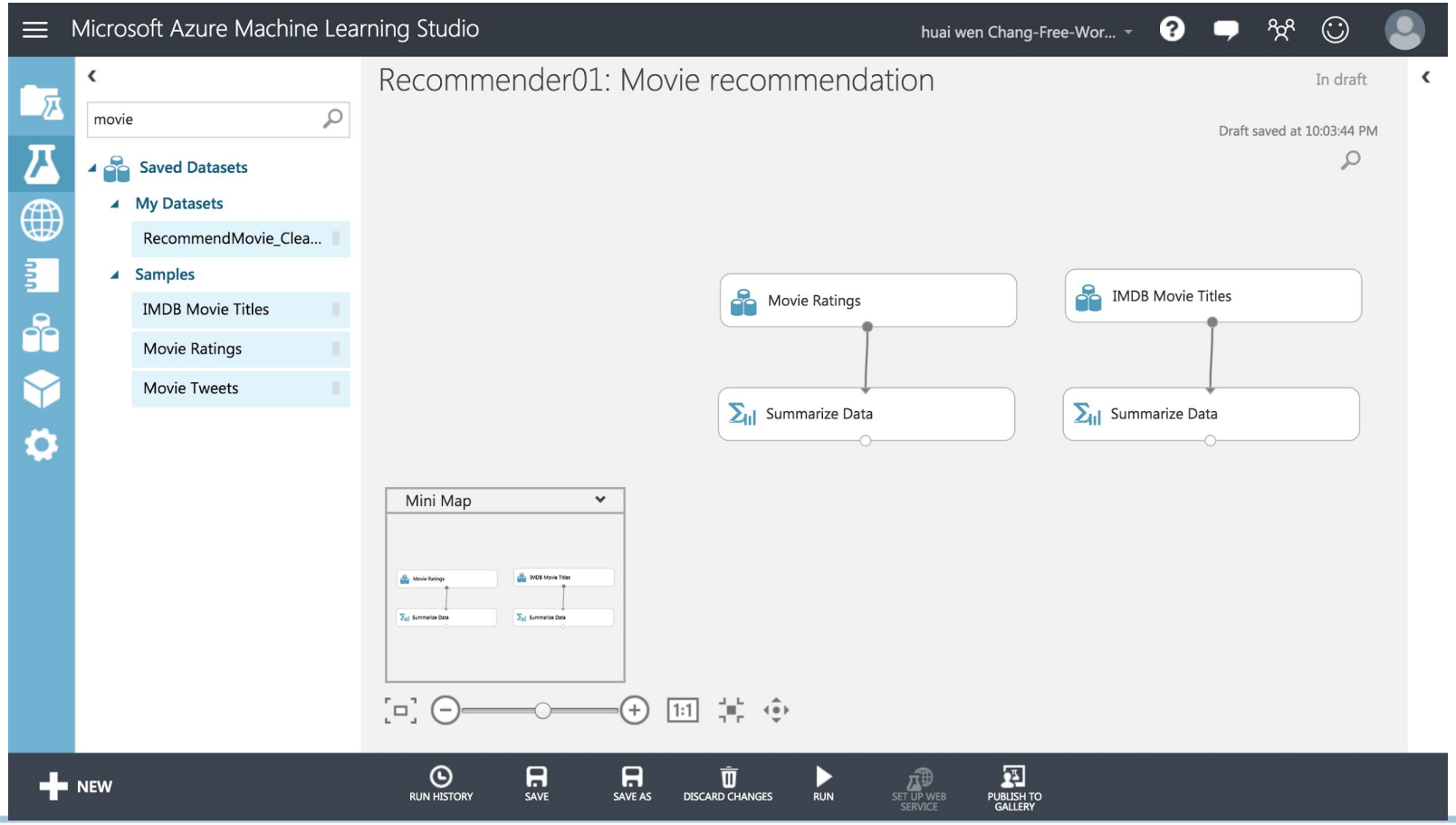
https://studio.azureml.net/



Experience Workflow



Checkpoint 01 - Viz Your Data



Checkpoint 01 - Viz Your Data

Recommender01: Movie recommendation › Movie Ratings › dataset

rows 227472 columns 4

view as

UserId	MovieId	Rating	Timestamp
1	68646	10	1381620027
1	113277	10	1379466669
2	454876	8	1394818630
2	790636	7	1389963947
2	816711	8	1379963769
2	1091191	7	1391173869
2	1322269	7	1391529691
2	1433811	8	1380453043

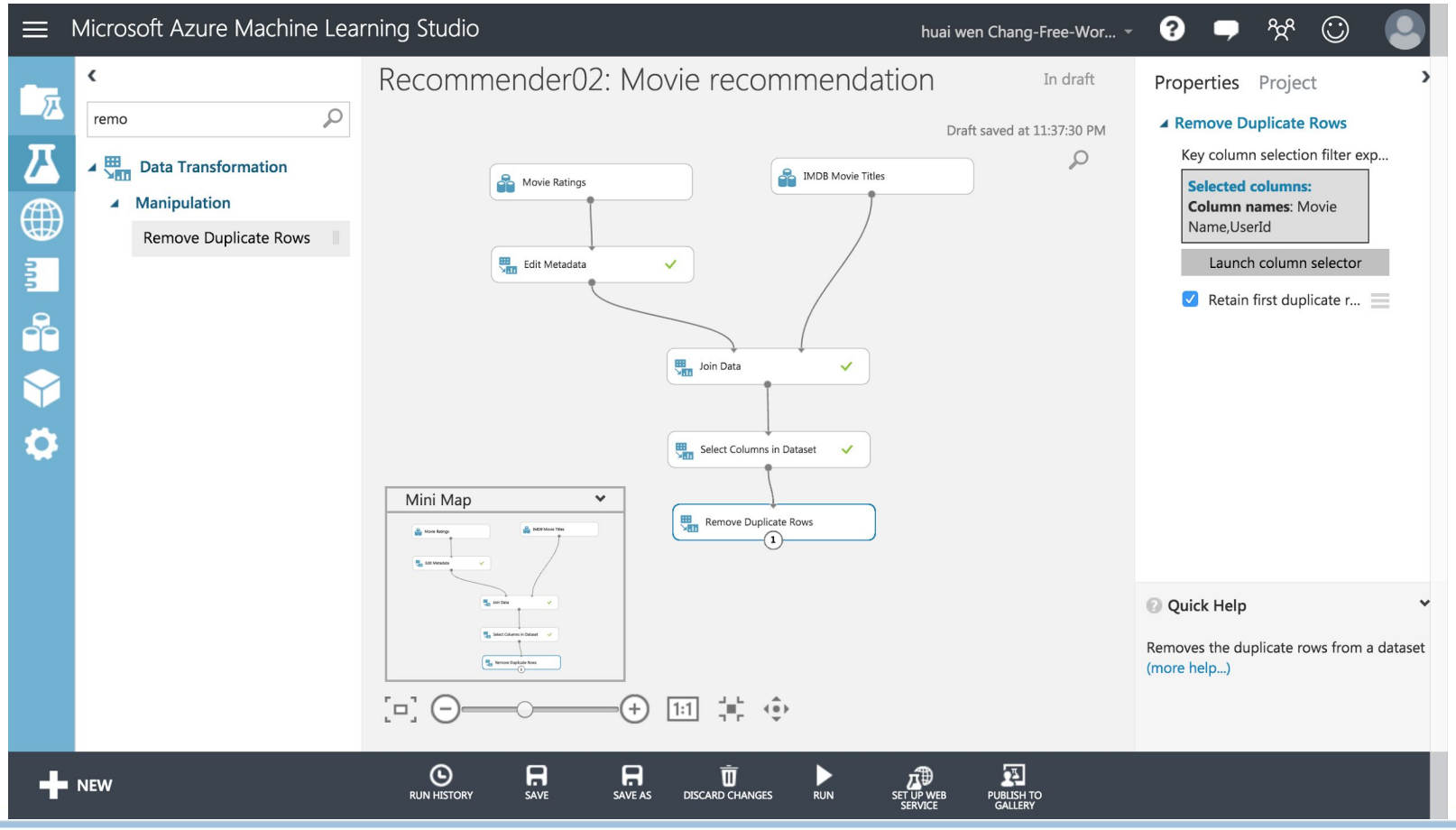
Recommender01: Movie recommendation › IMDB Movie Titles ›

rows 16614 columns 2

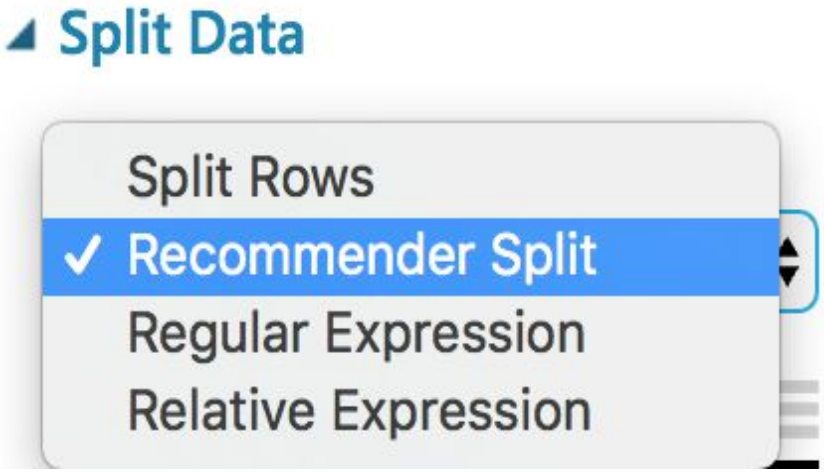
view as

Movie ID	Movie Name
8	Edison Kinetoscopic Record of a Sneeze (1894)
91	Le manoir du diable (1896)
417	Le voyage dans la lune (1902)
628	The s of Dollie (1908)
833	The Country Doctor (1909)
1223	Frankenstein (1910)
1740	The Lonedale Operator (1911)

Checkpoint 02 - Clean Your Data



Checkpoint 02 - Clean Your Data



Split Rows

> 一般分開資料的方法，預設是 50-50。

Recommender Split

> 推薦系統時會搭配的分資料方式，
確認user-item pairs 或是ratings 是被均勻分開的。

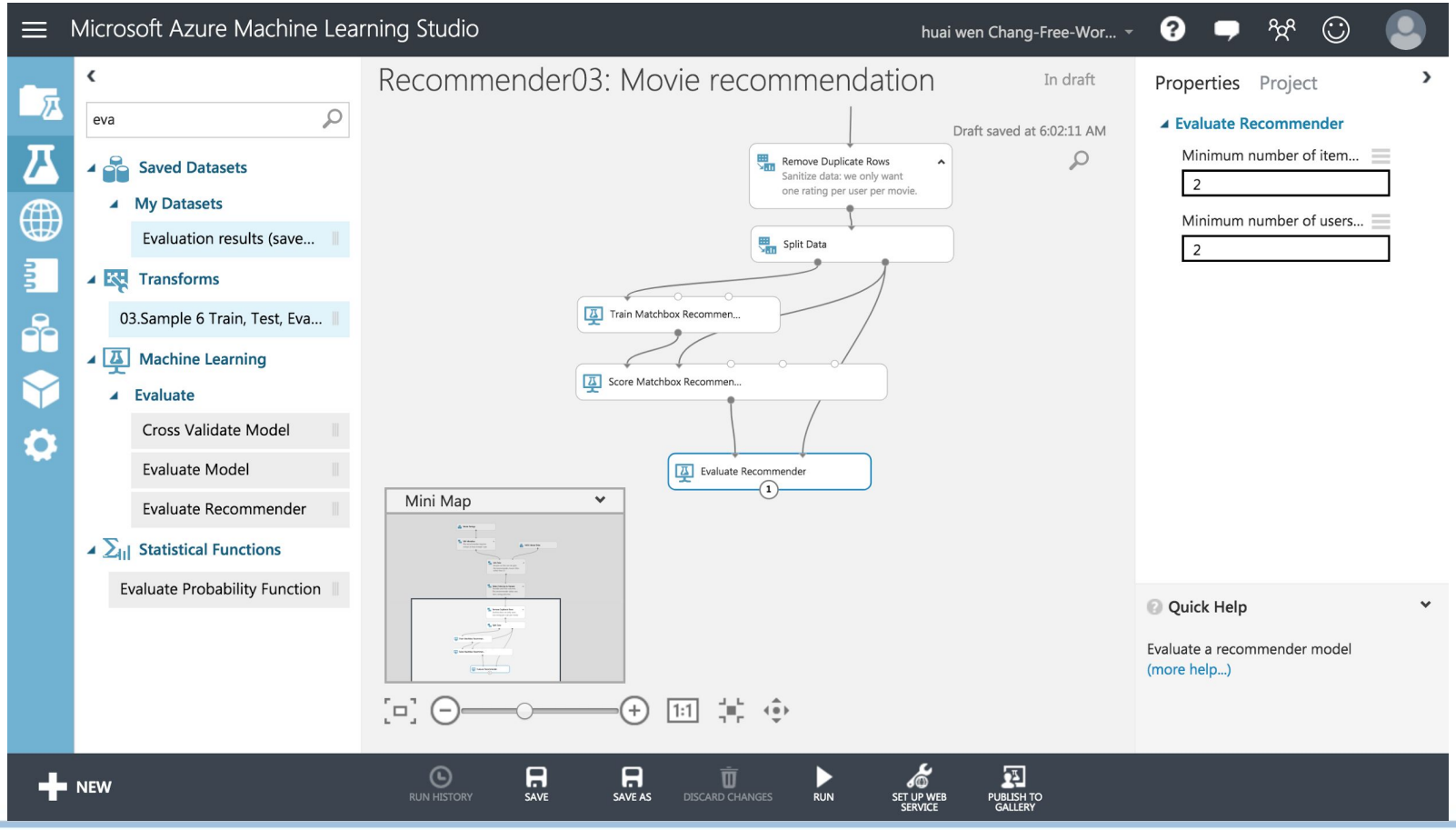
Regular Expression

> 文本情緒分析時使用

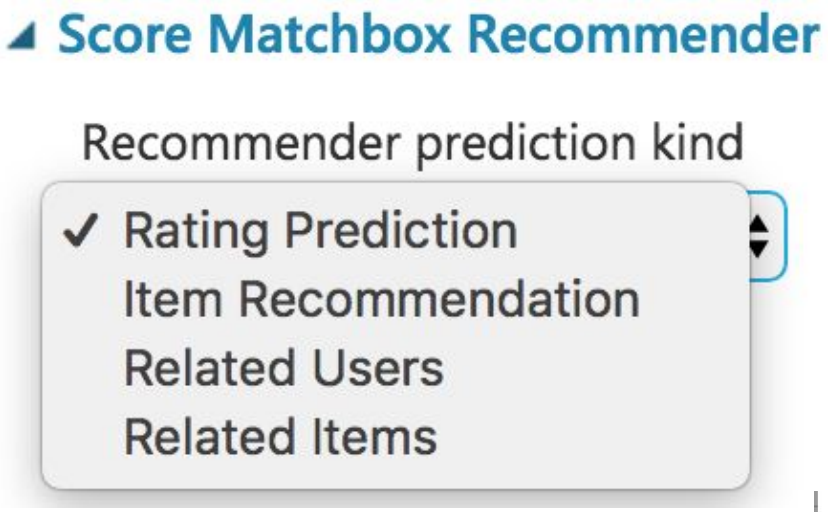
Relative Expression

> 使用時間，年紀，金額，或百分比當作是分開資料的依據時，
就可以選用這個方式

Checkpoint 03 - Training Model



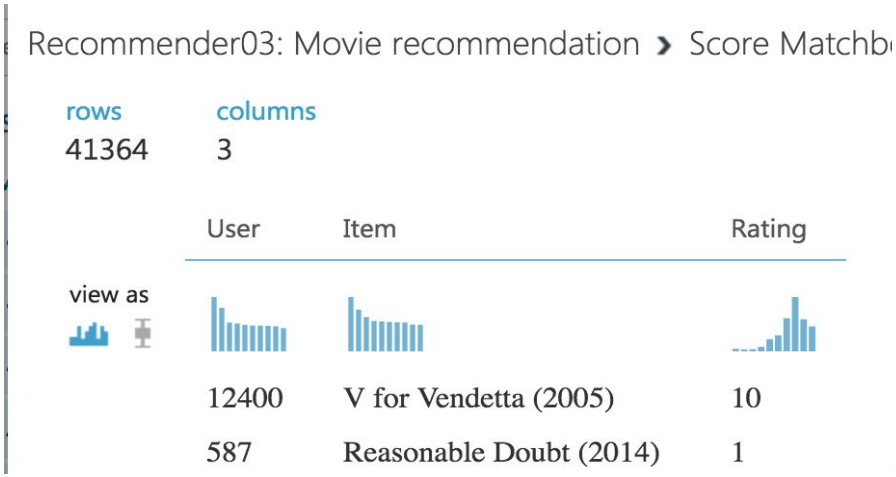
Checkpoint 03 - Training Model



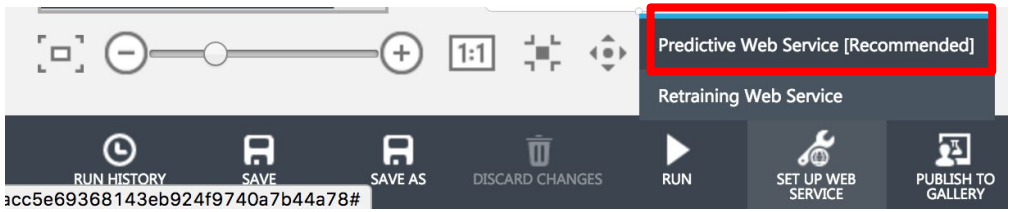
選擇想要Score的方式

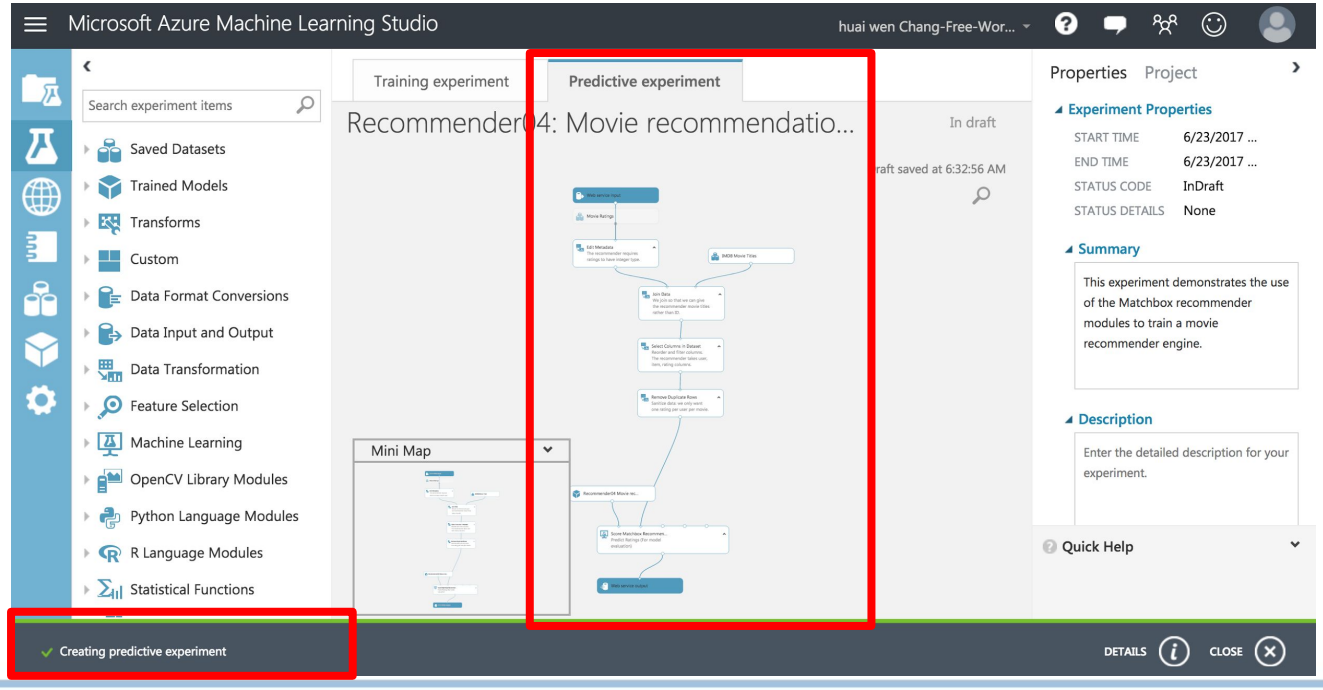
- Predict ratings for a given user and item
- Recommend items to a given user
- Find users related to a given user
- Find items related to a given item

依據不同方式放進input
輸出的結果(Visualize)

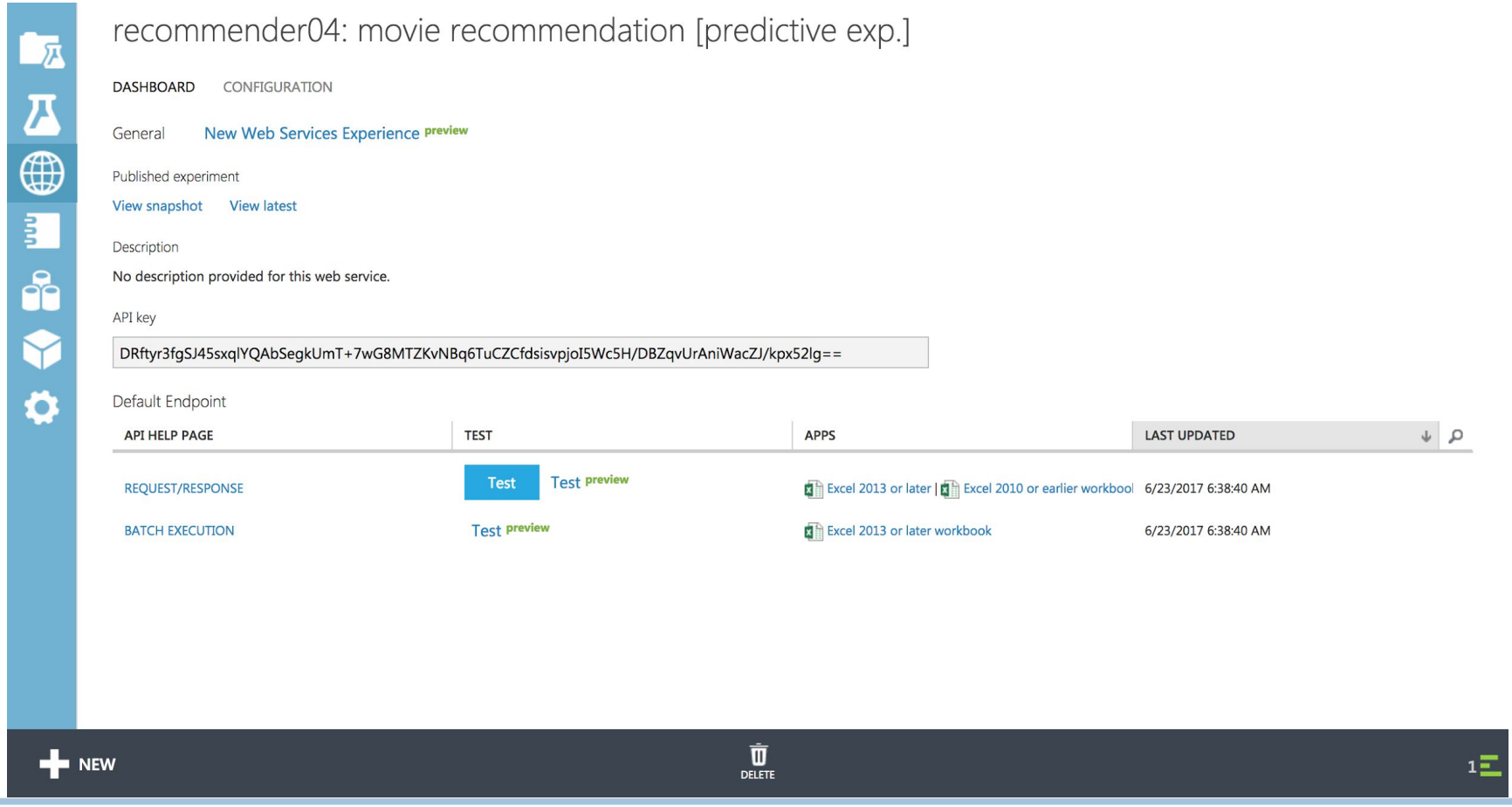


Checkpoint 04 - Evaluate & Publish





Checkpoint 04 - Evaluate & Publish





Q&A

對了，前面提到的都市傳說呢？



ref: 啤酒尿布是都市傳說

>> 是洗手乳


Dettol
滴露潔手液
促銷價 67元
滴露潔身液
自行走動者
來復易
長時間安心
長期臥床者適用
可自行走動者
防漏安心
輕薄安心

Further Reading



ref: What the Million-Dollar Netflix Prize Taught Me About Marketing

Read More

- 實驗

  AzureML Recommender: Movie recommendation

- 演算法論文

  MS Research Matchbox: Large Scale Bayesian Recommendations

- 推薦系統背景知識

  MS Technet Recommendations Everywhere

Contact

facebook.com/mia5419

twMVC

R-Ladies

Azure Taiwan

Tech Podcast Night



Blog 是記錄知識的最佳平台



感謝 Jetbrains 贊助贈品



ReSharper

https://www.jetbrains.com/resharper/

感謝 OzCode 贊助贈品



http://www.oz-code.com/

業界師資、實戰教學



http://skilltree.my

謝謝各位

http://mvc.tw

- 本投影片所包含的商標與文字皆屬原著作者所有。
- 本投影片使用的圖片皆從網路搜尋。
- 本著作係採用姓名標示 -非商業性-相同方式分享 3.0 台灣授權。閱讀本授權條款，請到 http://creativecommons.org/licenses/by-nc-sa/3.0/tw/，或寫信至Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.