資料科學研究院

挖掘出大數據的隱藏寶藏-Azure Machine Learning

Allan Yiin

asiaMiner

CURVE UP THE FUTURE

Agenda

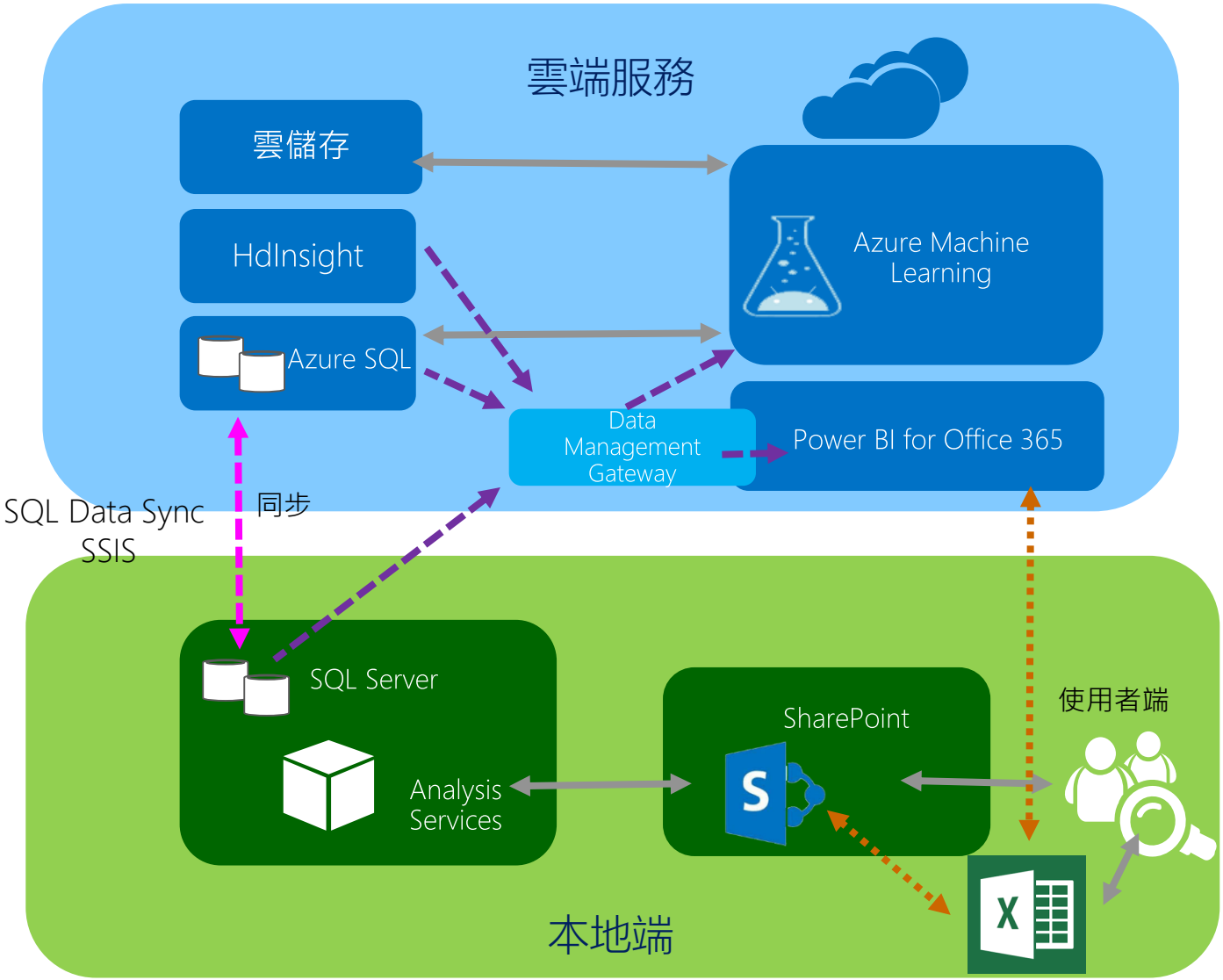
- 微軟雲端大數據分析架構
- Azure Machine Learning架構介紹
- 主要功能介紹
- 案例：建置回應模型
- 案例：客戶區隔模型
- 整合R語言
- 整合Power BI for Office 365

雲端大數據分析架構

大數據時代的到來


ExaByte
（10E18）
PetaByte
（10E15）
TeraByte
（10E12）
GigaByte
（10E9）
大數據
物聯網
社會輿論
點擊流
感測器/ 射頻識別 / 設備
維基 / 微博
音訊 / 視頻
系統日誌
空間 & GPS 定位座標
整合資料市場
整合e化政府
天氣
文字/圖片
WWW 2.0
手機
廣告
電子商務
合作
數位行銷
搜索行銷
網路日誌
推薦
企業資源規劃 / 客戶關係管理
應收款管理
聯繫管理
薪酬管理
交易追蹤
庫存管理
銷售管道
快速增長
多變
多樣性
企業資源規劃 / 客戶關係管理
WWW 2.0
物聯網
1980
190,000$
1990
9,000$
2000
15$
2010
0.07$

雲端與本地端的分析架構



有了大數據之後
如何從中產生效益...?

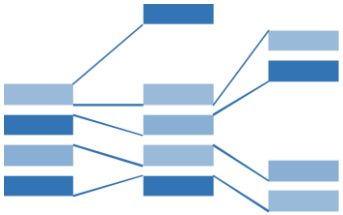
微軟Cortana世界盃足球賽16場正確預測15場比賽...
擁有數據，就能預測這世界的未來...

Azure Machine Learning

Machine Learning在做些甚麼...?

透過演算法協助我們從龐大的數據中找出有意義的規則

分類



分群

推估

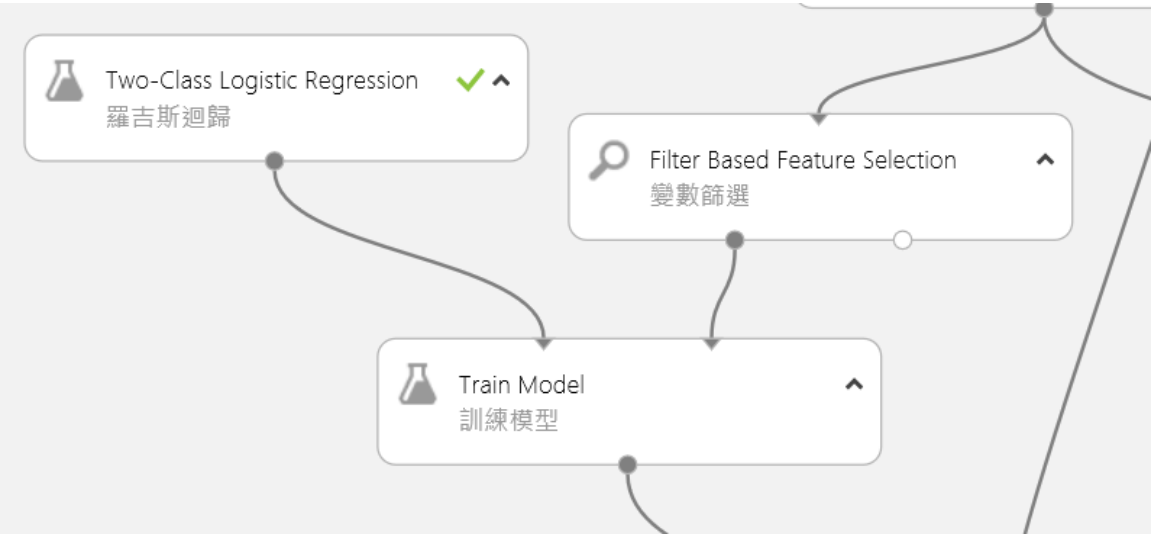
Azure Machine Learning

透過瀏覽器存取，將分析功能透過流程圖的模式進行設計...

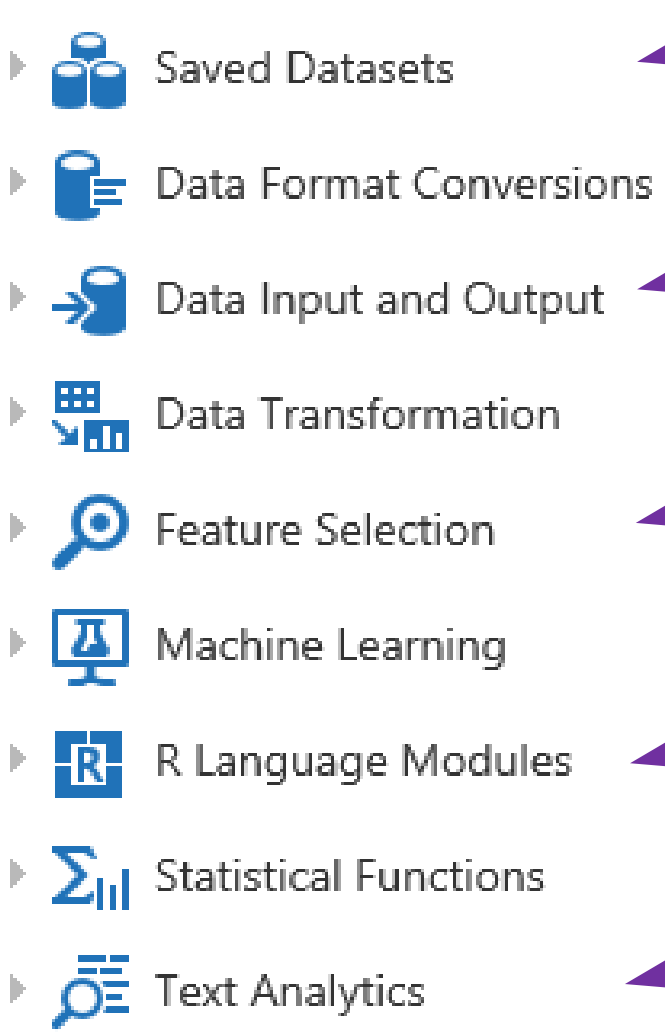
免費版本：
https://studio.azureml.net/

Azure Machine Learning包括了...

- 資料源的存取(與Azure SQL Database, Azure Storage以及HDInsight深度整合)
- 資料清理(抽樣、樣本切割、計算欄位...)
- 敘述性統計
- 有效變數選取
- 建立預測模型
- 評估預測模型
- 應用預測模型



主要功能介紹



儲存資料集，存放於Azure Storage的檔案

檔案結構轉檔

雲端資料的寫入與寫出

資料預處理與操作

變量選取

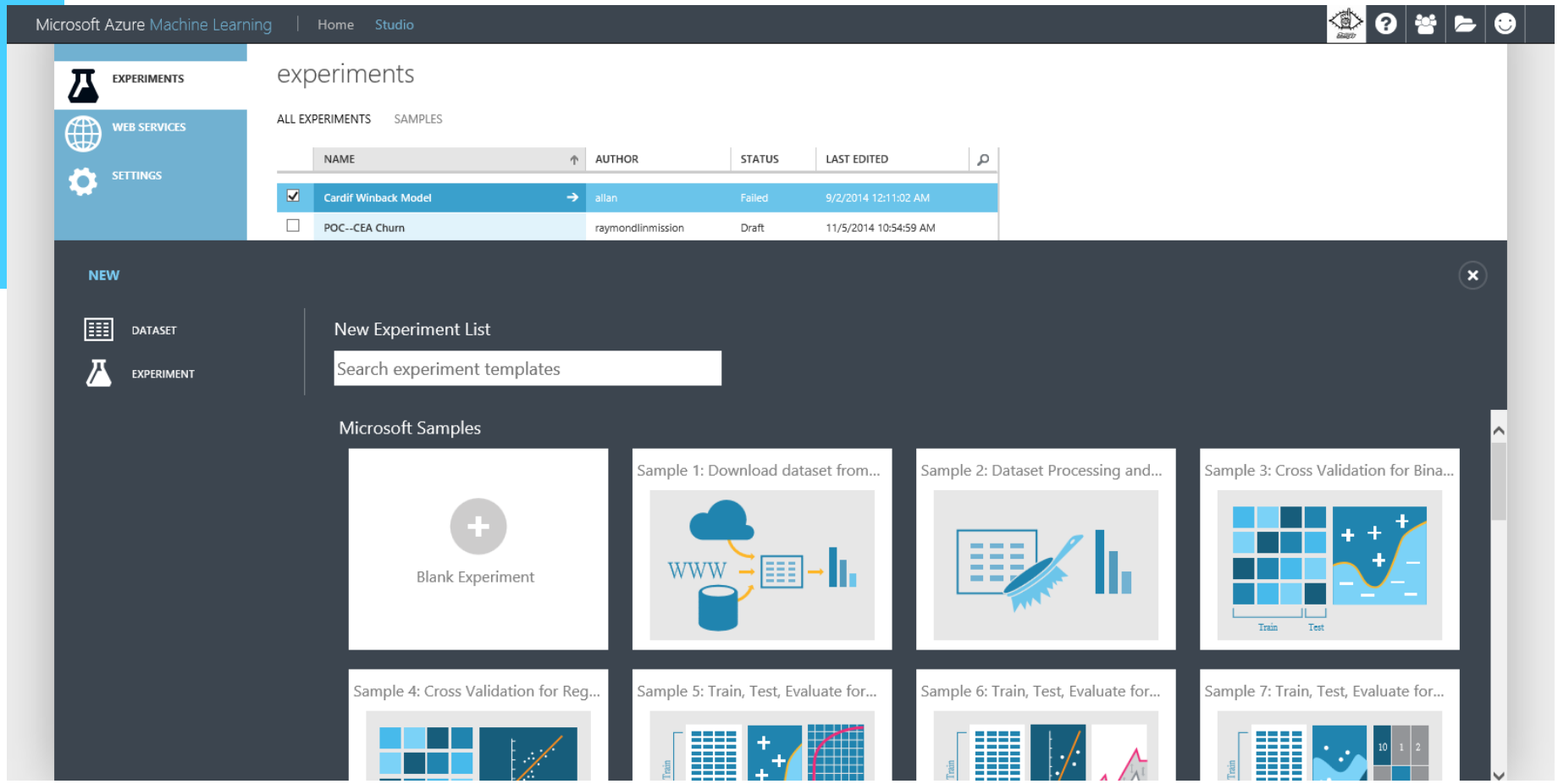
演算法、訓練、評估模型以及進行預測

使用R語言

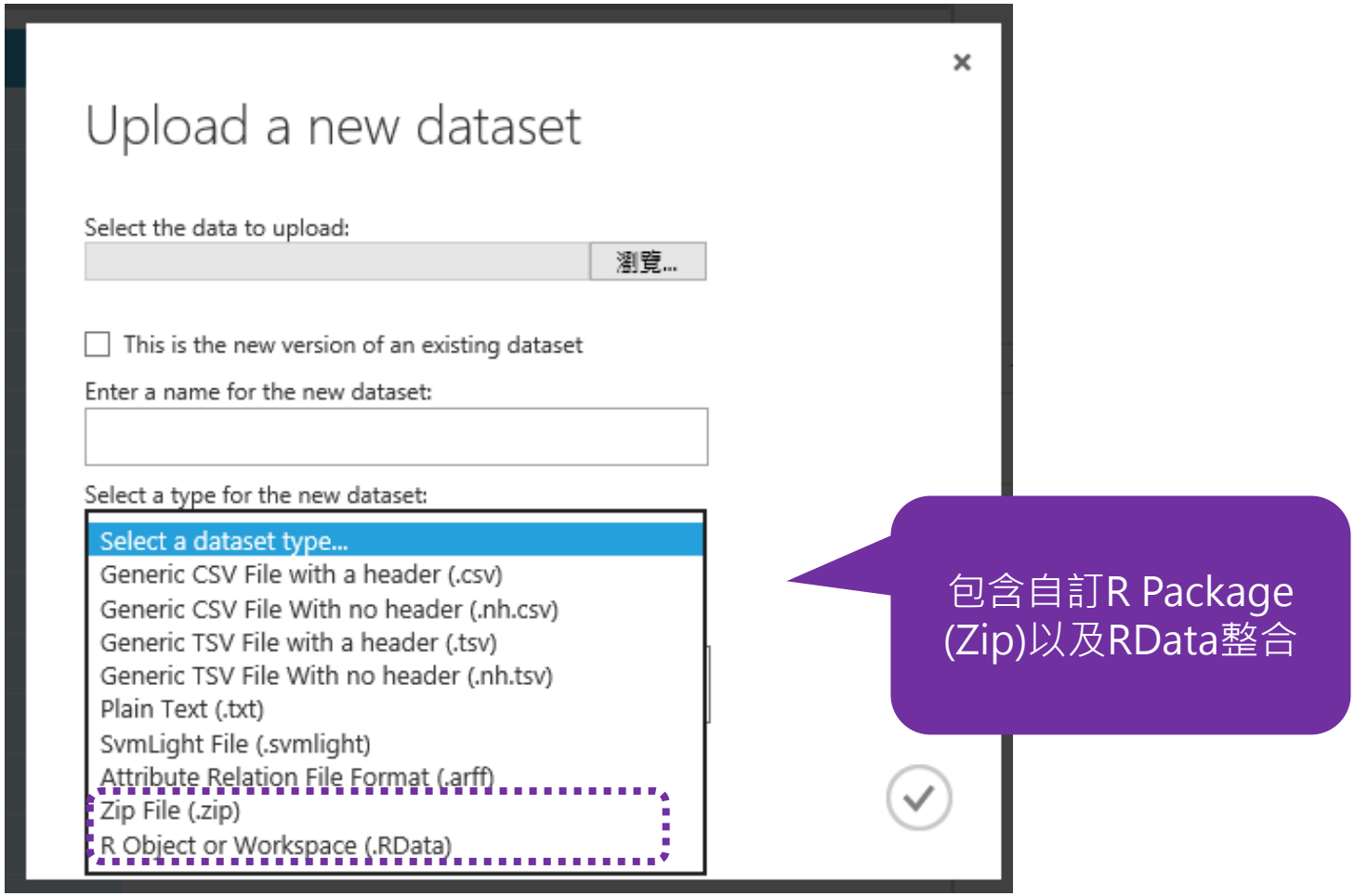
統計分析與敘述性統計

文字資料採礦(限英文)

新增資料集或Experiment

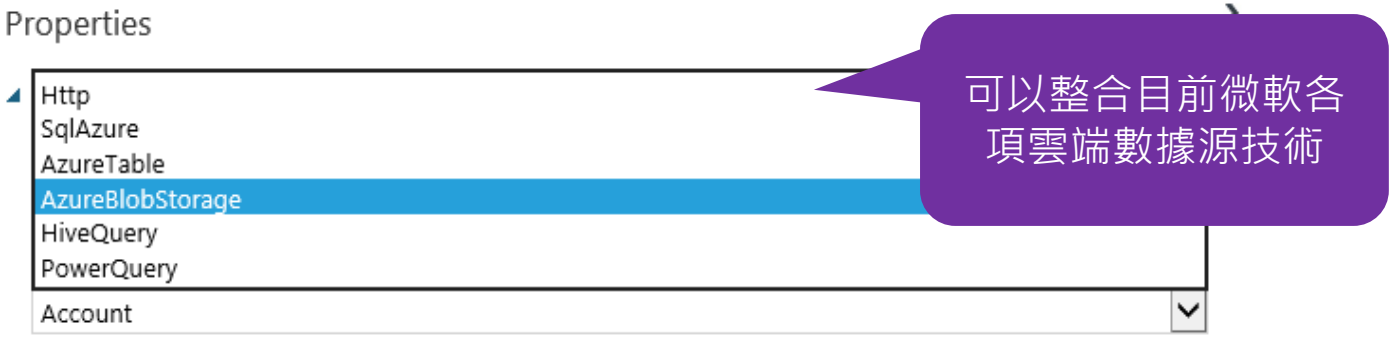


上傳資料集

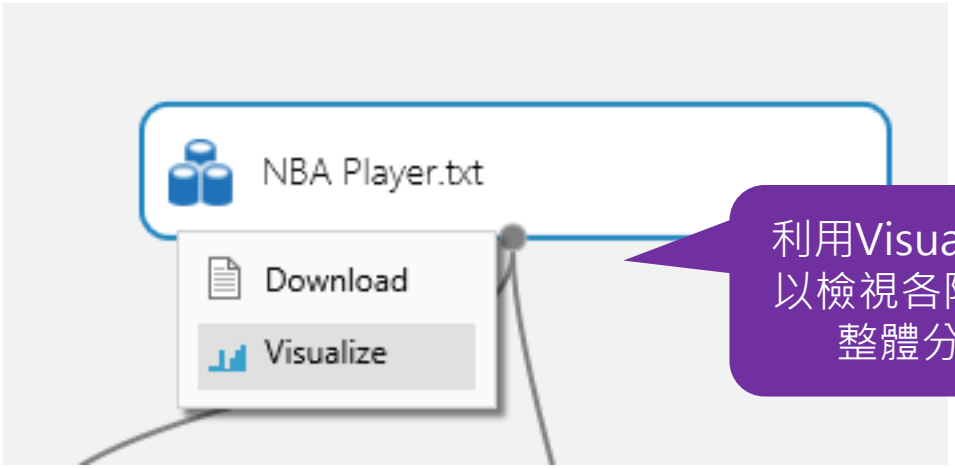

Upload a new dataset
Select the data to upload:
瀏覽...
This is the new version of an existing dataset
Enter a name for the new dataset:
Select a type for the new dataset:
Select a dataset type...
Generic CSV File with a header (.csv)
Generic CSV File With no header (.nh.csv)
Generic TSV File with a header (.tsv)
Generic TSV File With no header (.nh.tsv)
Plain Text (.txt)
SvmLight File (.svmlight)
Attribute Relation File Format (.arff)
Zip File (.zip)
R Object or Workspace (.RData)
包含自訂R Package (Zip)以及RData整合

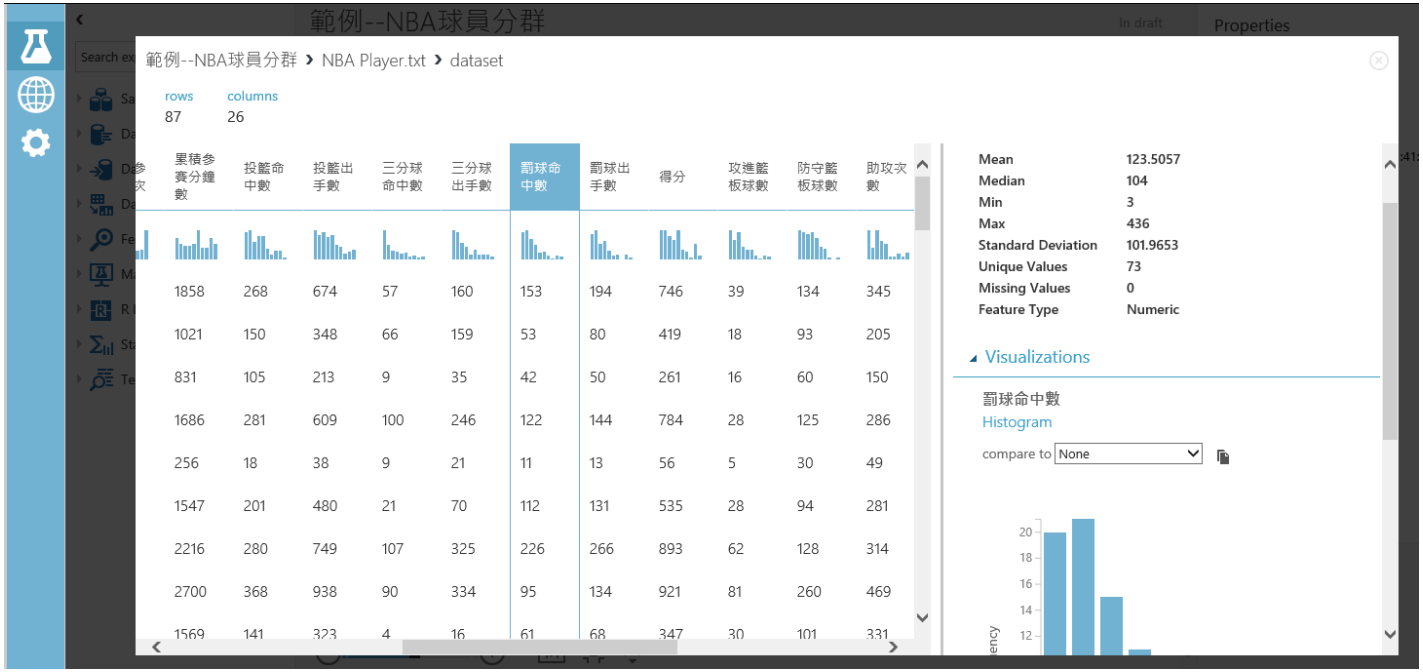
雲端資料讀取


Reader


Properties
Http
SqlAzure
AzureTable
AzureBlobStorage
HiveQuery
PowerQuery
Account
可以整合目前微軟各項雲端數據源技術

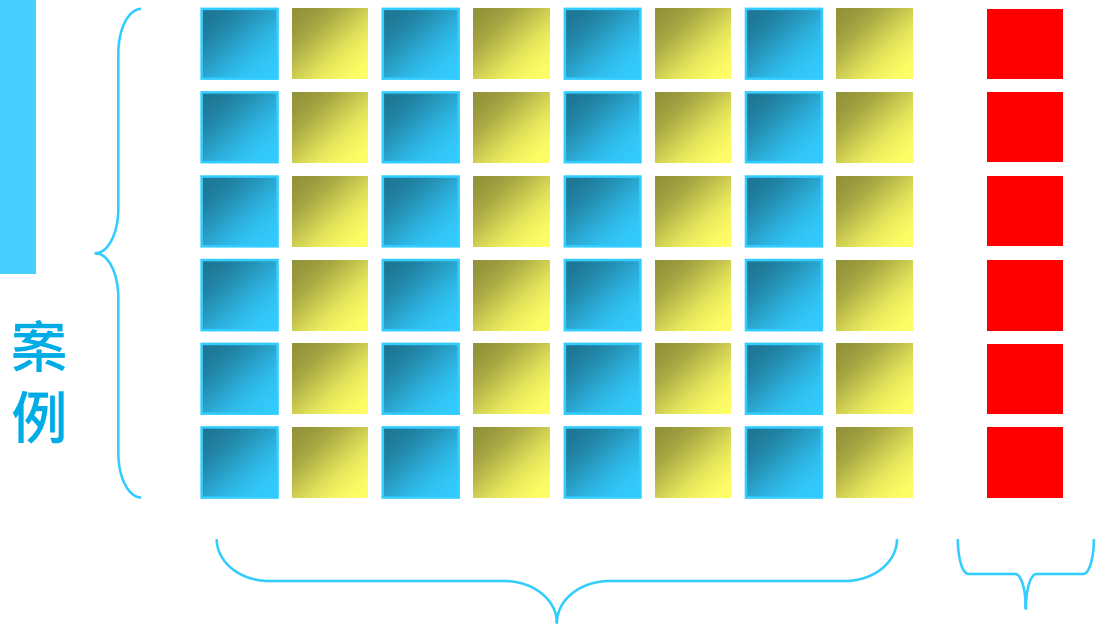
變數敘述性統計


NBA Player.txt
Download
Visualize
利用Visualize功能可以檢視各階段數據集整體分布狀況


範例--NBA球員分群
In draft
Properties
範例--NBA球員分群 › NBA Player.txt › dataset
rows 87
columns 26
Mean 123.5057
Median 104
Min 3
Max 436
Standard Deviation 101.9653
Unique Values 73
Missing Values 0
Feature Type Numeric
Visualizations
罰球命中數
Histogram
compare to None

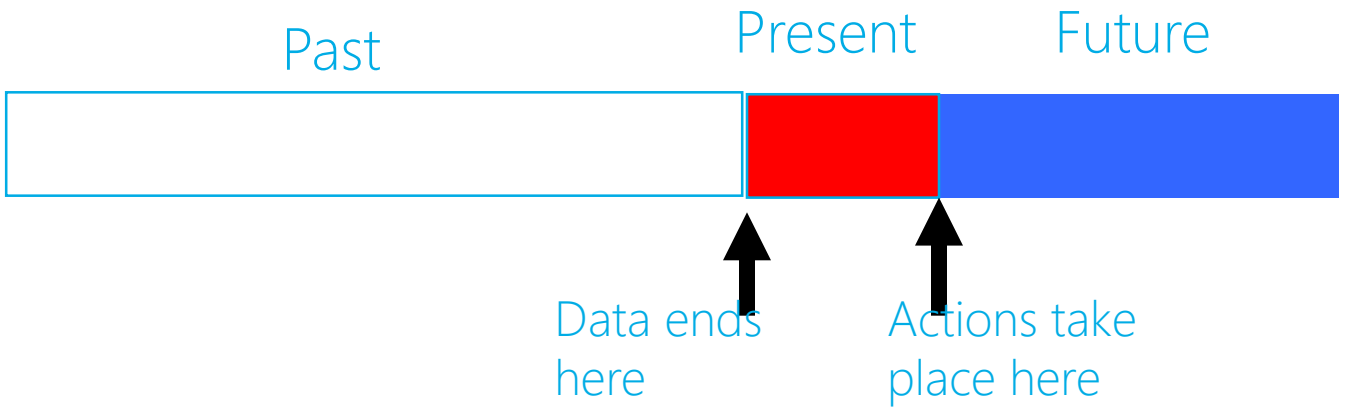
建置
回應模型

什麼是預測？



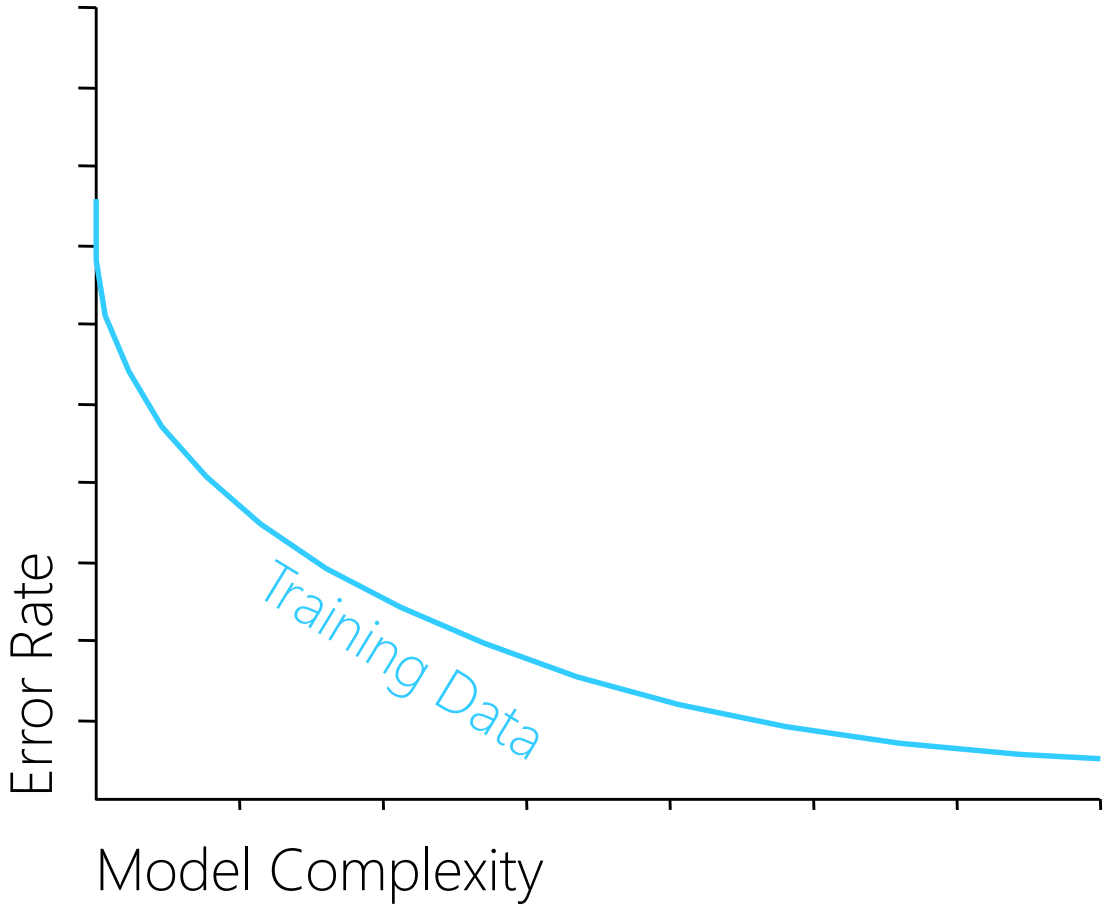
- 預測的前提：使用A預測B
  - → A必須發生於B之前
  - → 在B發生之前必須能取得A的資料

資料採礦是關於時間的

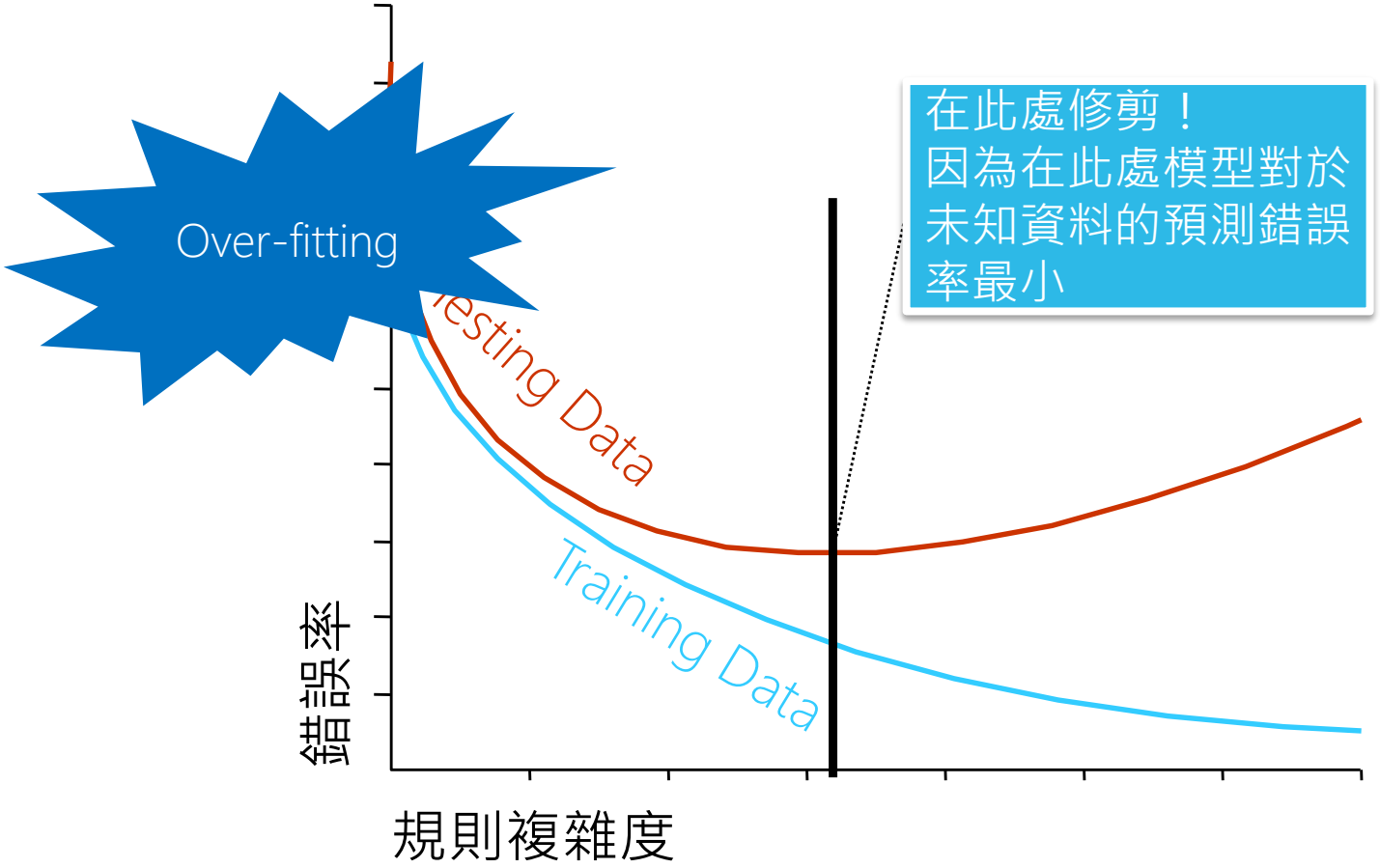


- 建立模型 **使用歷史資料**--結果是已知的
- 應用 (或是 評分) 模型 **使用現有的資料**
- 執行結果 **使用未來的資料**(下週或是下個月)--結果是未知的

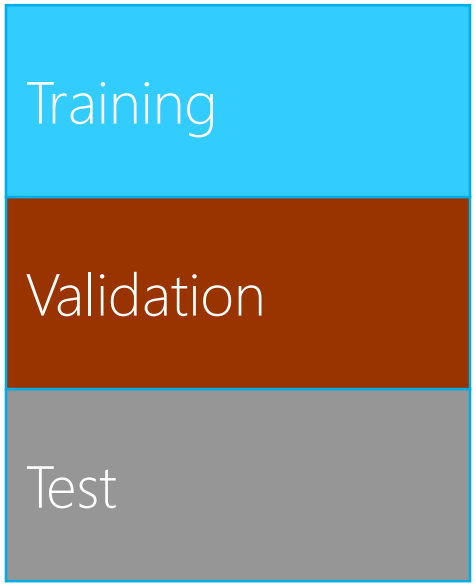
模型穩定性挑戰



模型測試集的結果


Over-fitting
在此處修剪！
因為在此處模型對於未知資料的預測錯誤率最小
esting Data
Training Data
錯誤率
規則複雜度

樣本切割

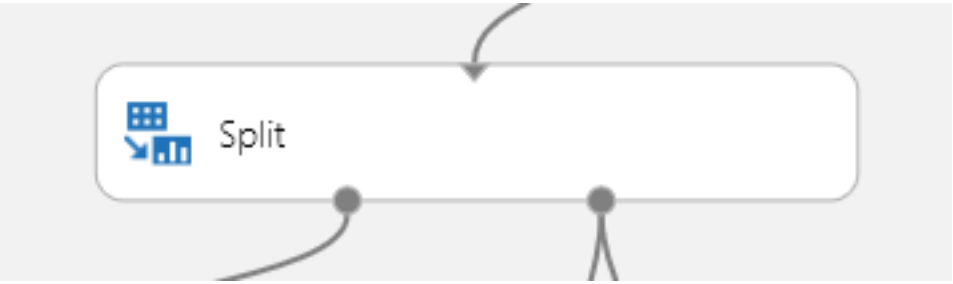


使用訓練組資料建立預測模型

使用鑑校組資料於建模階段評估模型

使用測試組資料驗證模型

樣本切割



Properties

Split

Splitting mode

Split Rows

Fraction of rows in the first output dataset

0.6

Randomized split

Random seed

0

Stratified split

False

切割資料列

切割比例

變數選取

Properties

Filter Based Feature Selection

Feature scoring method

Mutual Information

變量選取方法

Operate on feature columns only

Target column

Selected columns:
Column names: Response

設定預測變數

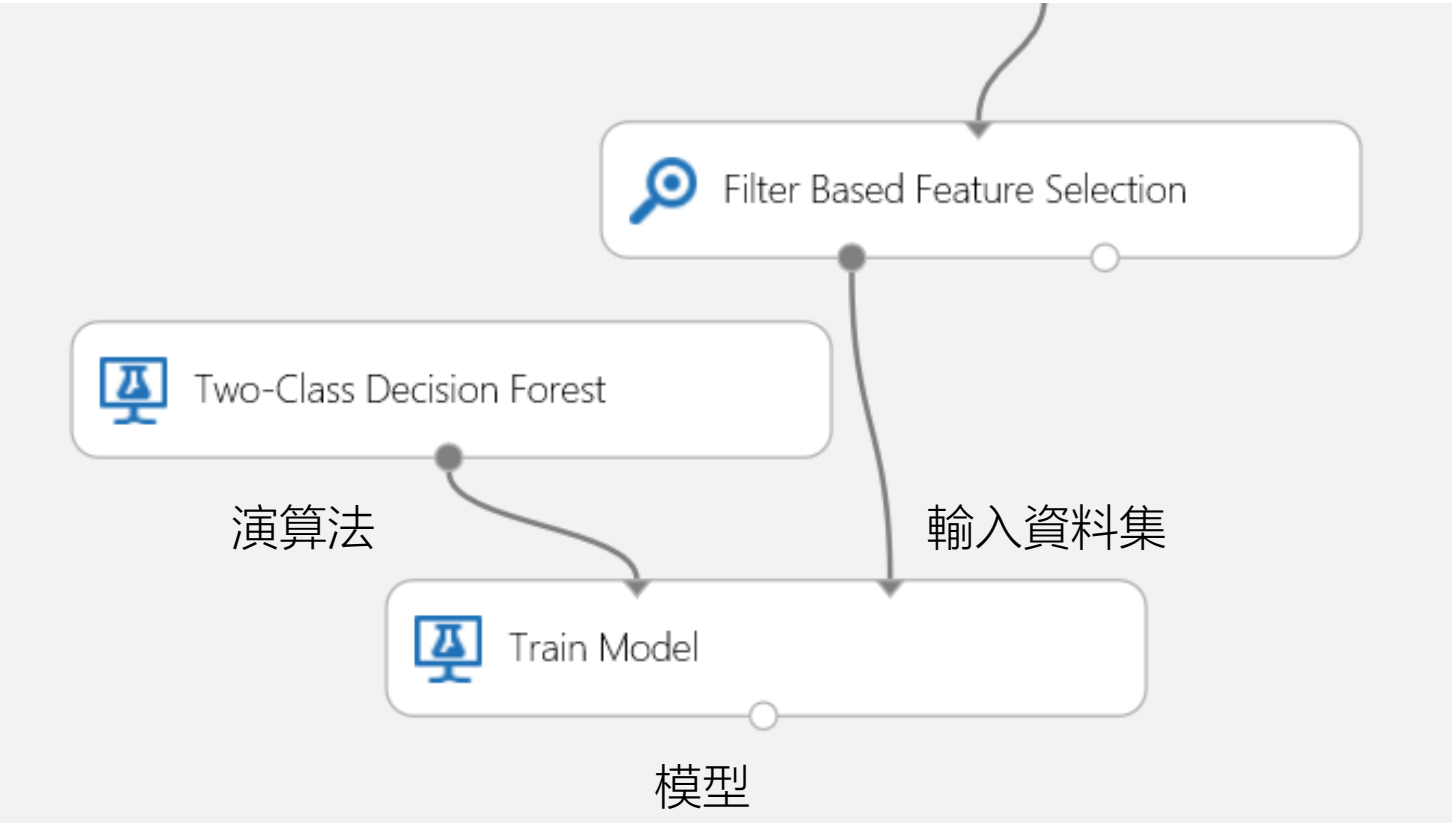
Launch column selector

Number of desired features

10

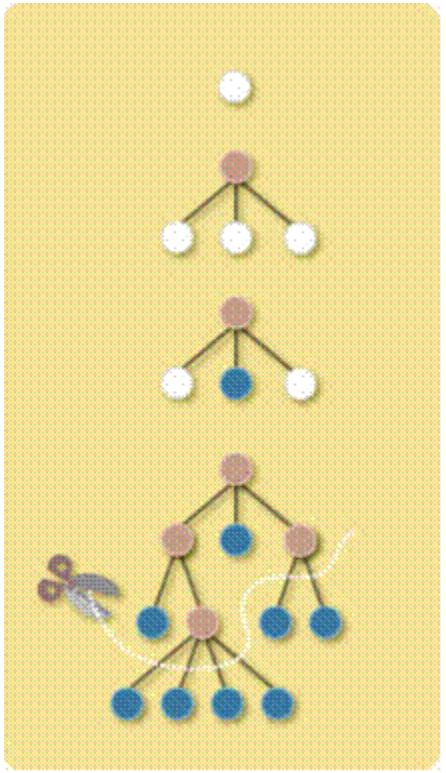
設定選入變數量

建立預測模型


Filter Based Feature Selection
Two-Class Decision Forest
演算法
輸入資料集
Train Model
模型

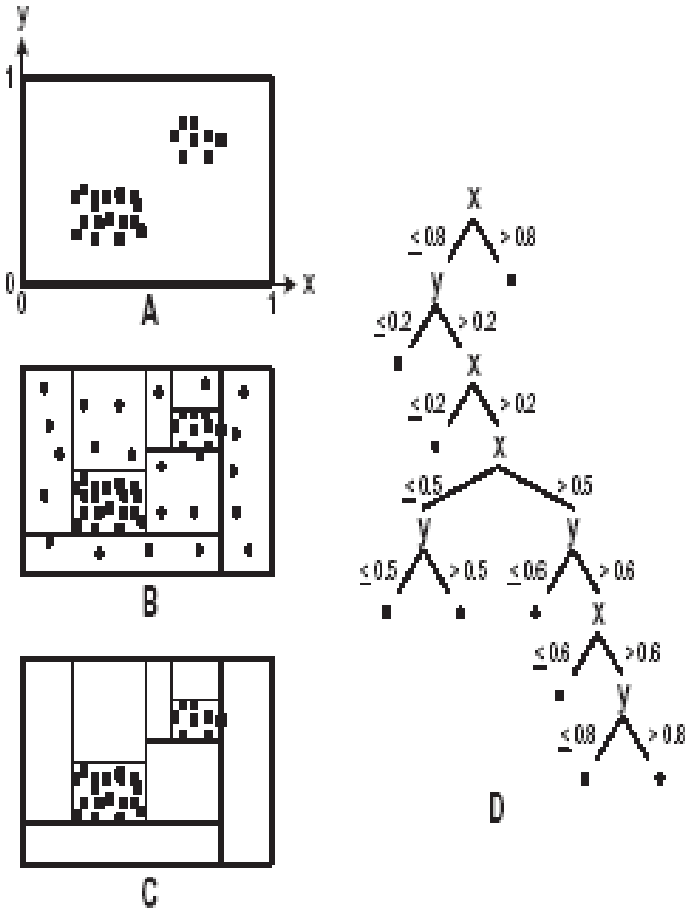
決策樹(Decision Tree)

- 透過樹狀規則呈現分類準則
- 找出最佳分岔點，使得資料的亂度最低
- 根據每個分岔的分布來決定預測機率

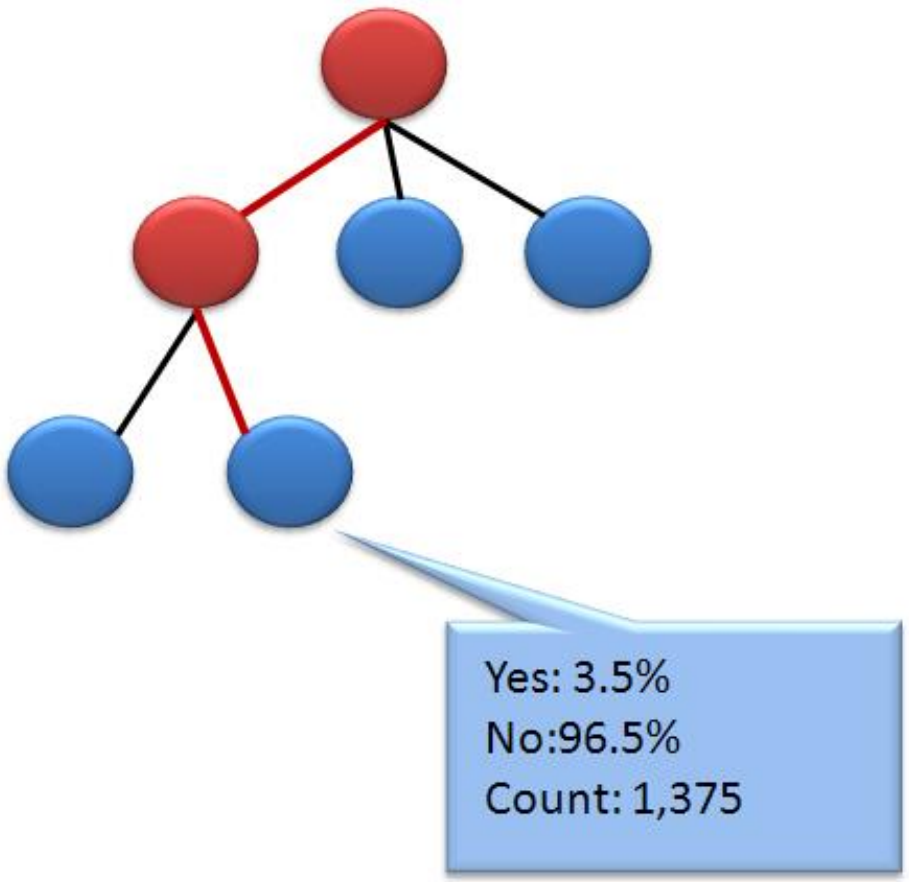


決策樹

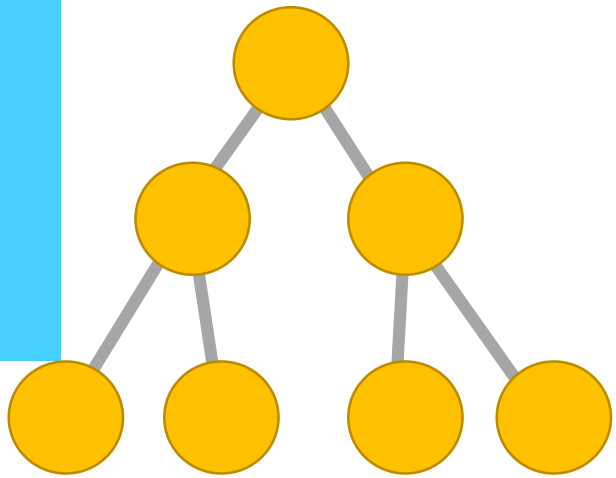
- 決策樹適合做為多維空間中方型切割以找尋最佳



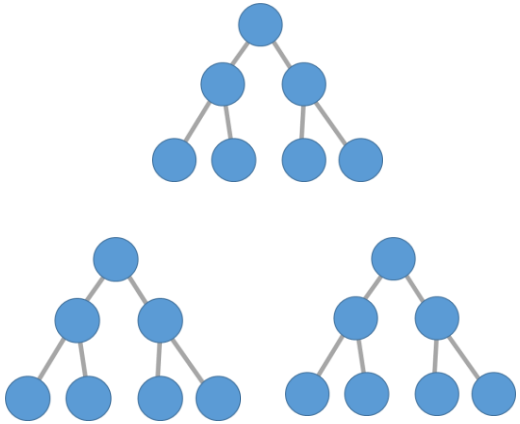
指派機率



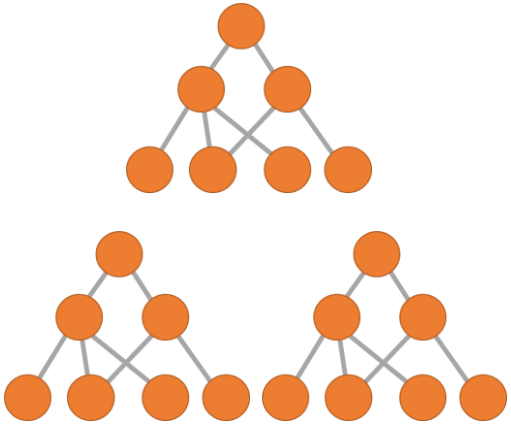
新型態的演算法



Decision Tree
決策樹



Decision Forest
決策森林

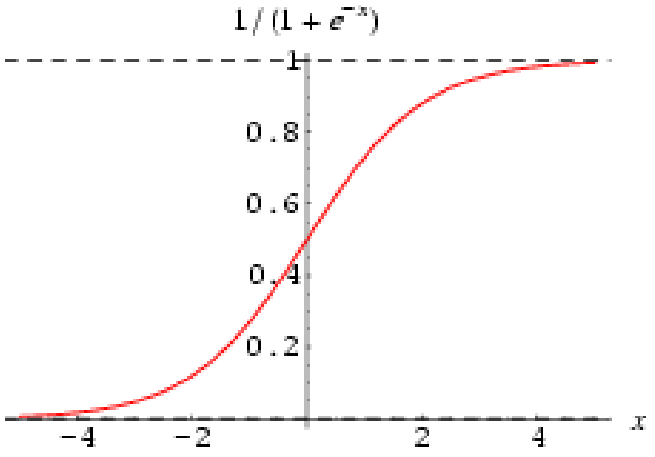


Decision Jungle
決策叢林

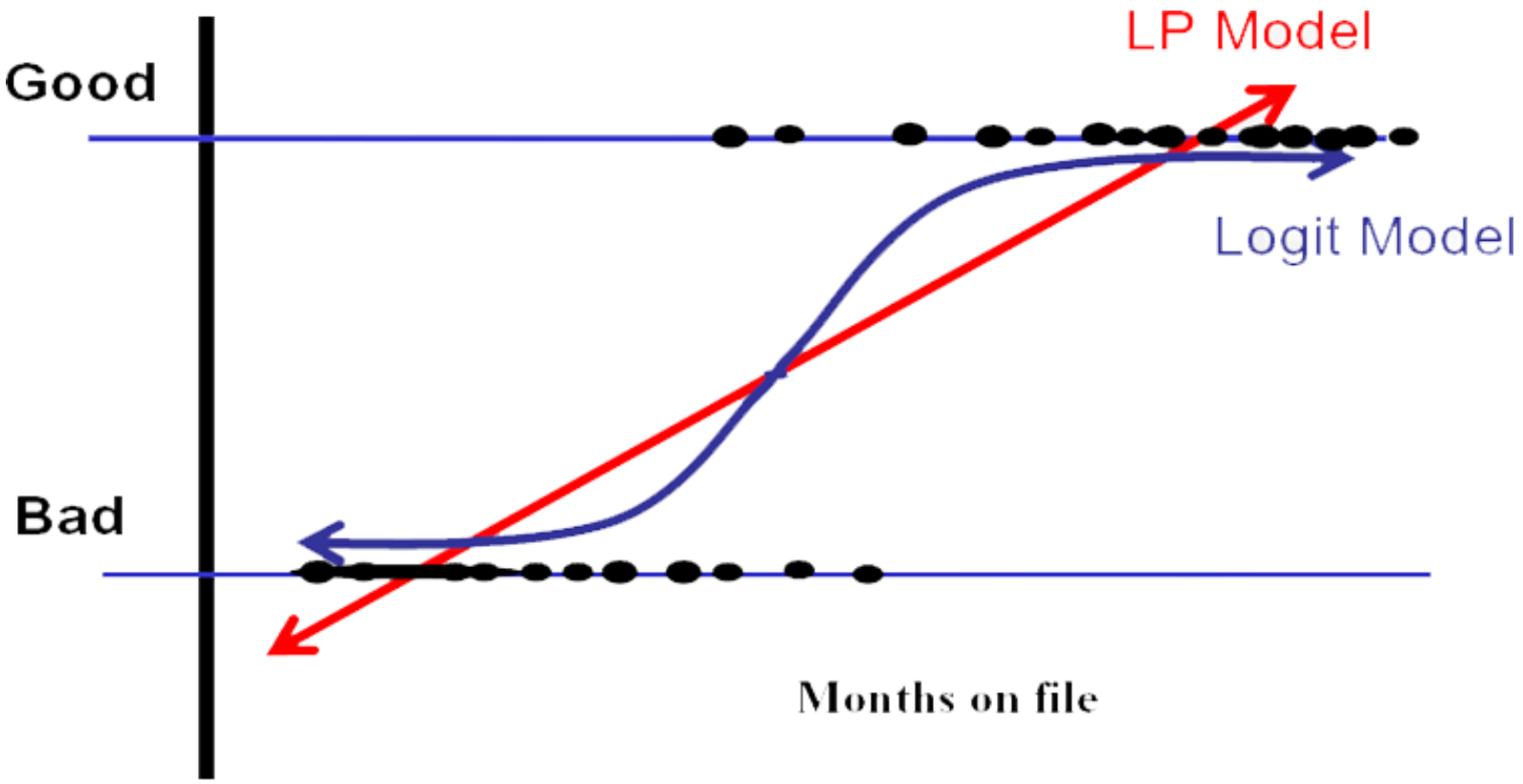
Logistic Regression

- Ln(P/1-P)=線性方程式
- P/(1-P)=odds rate
- 二元分類
- 輸入變數必須符合迴歸方程式，呈現單調遞增或是單調遞減趨勢

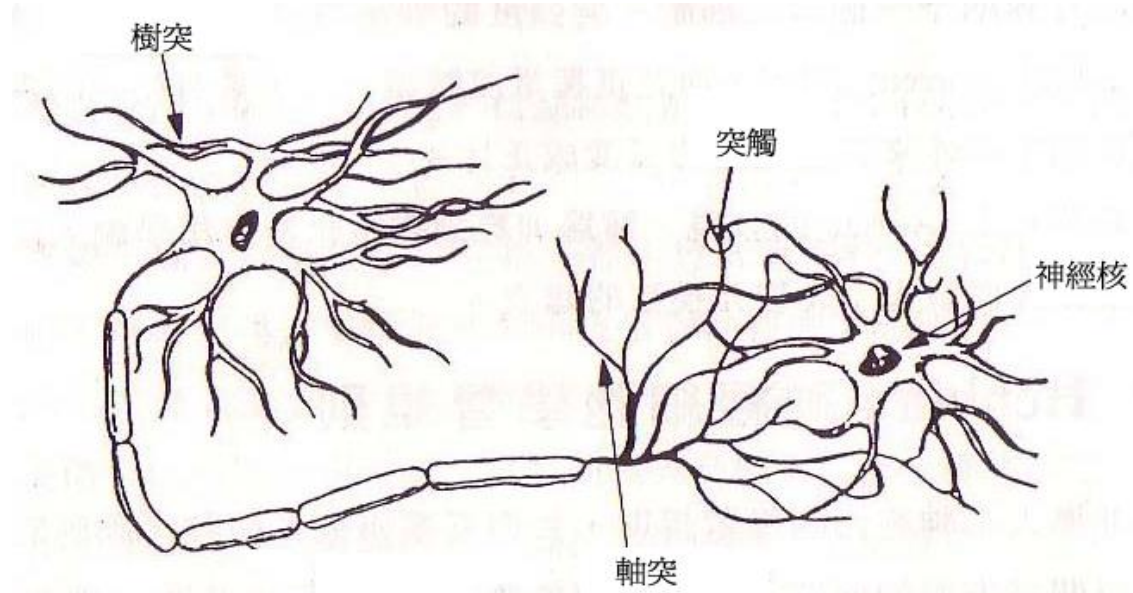
$$\ln\left[p/(1-p)\right] = a_0 + a_1 x_1 + \cdots + a_k x_k$$

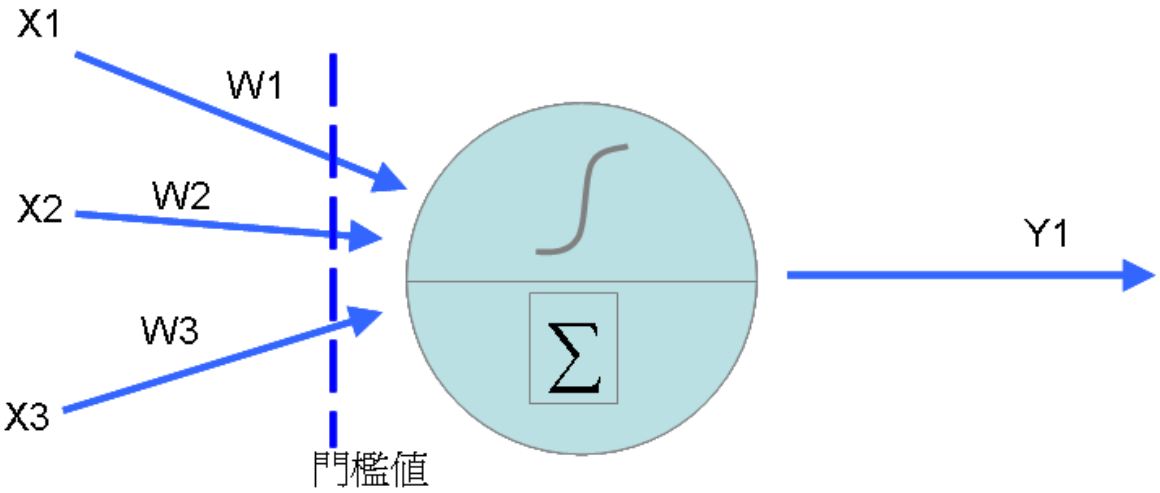


Logistic Regression

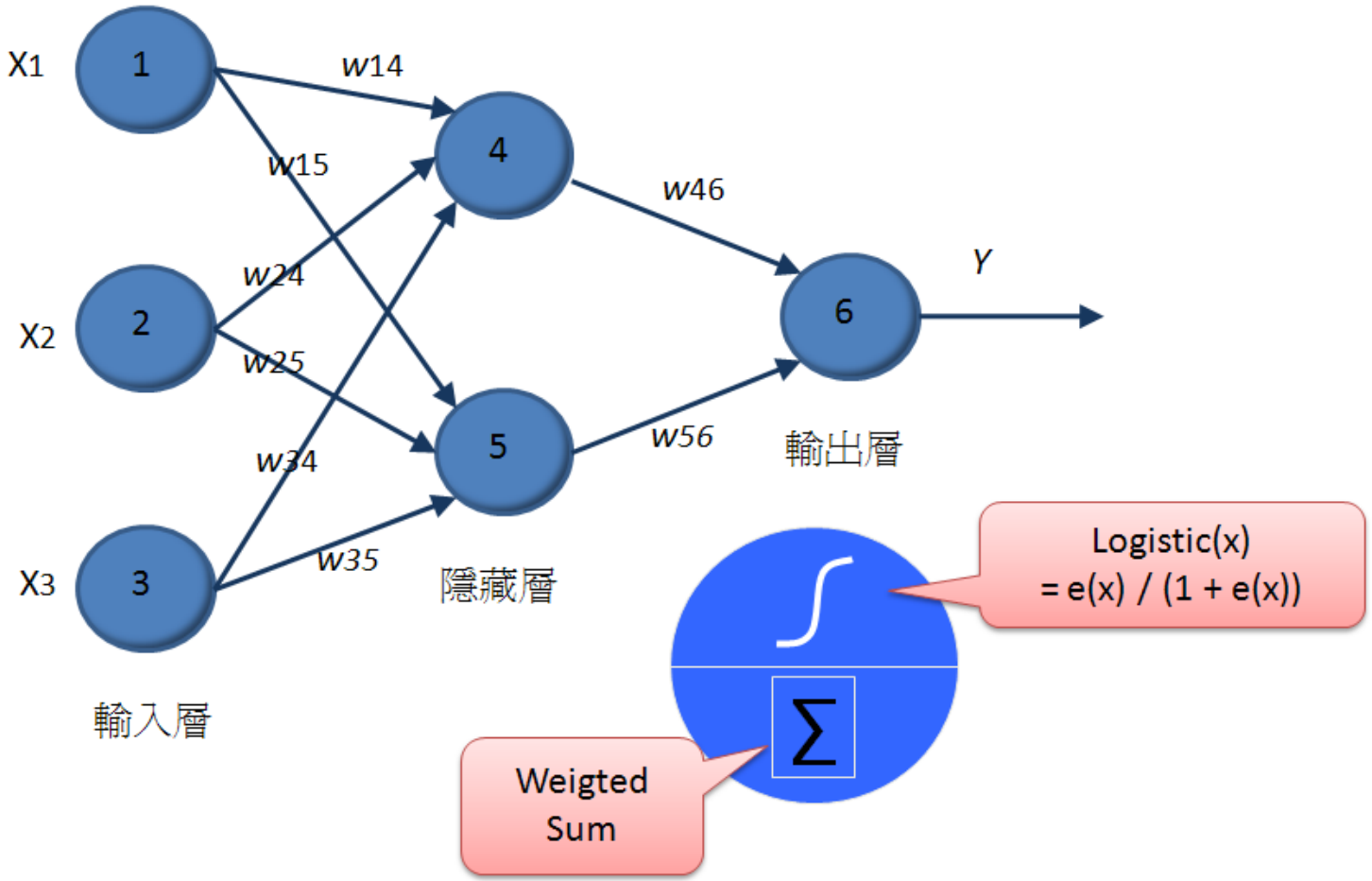


類神經網路

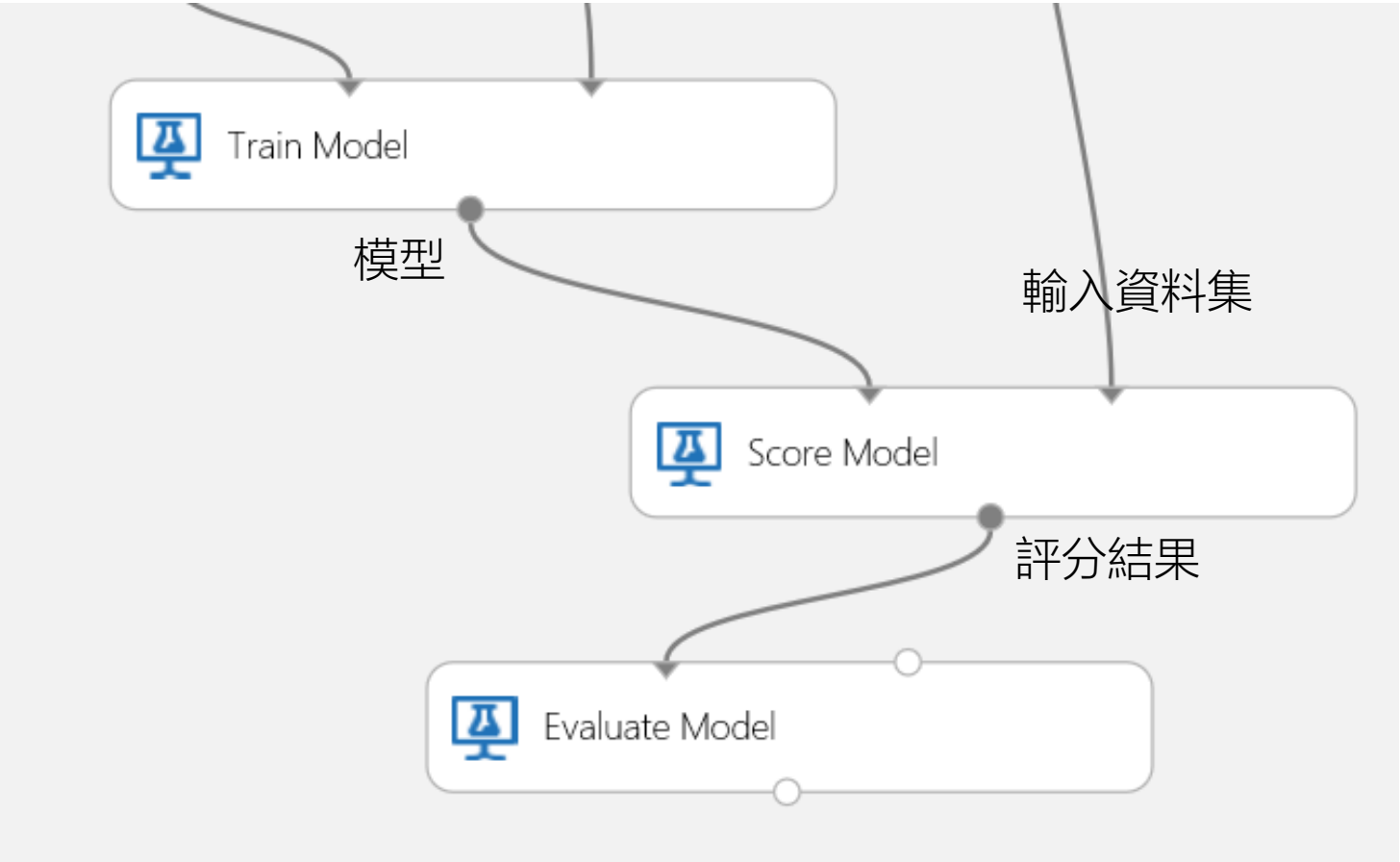




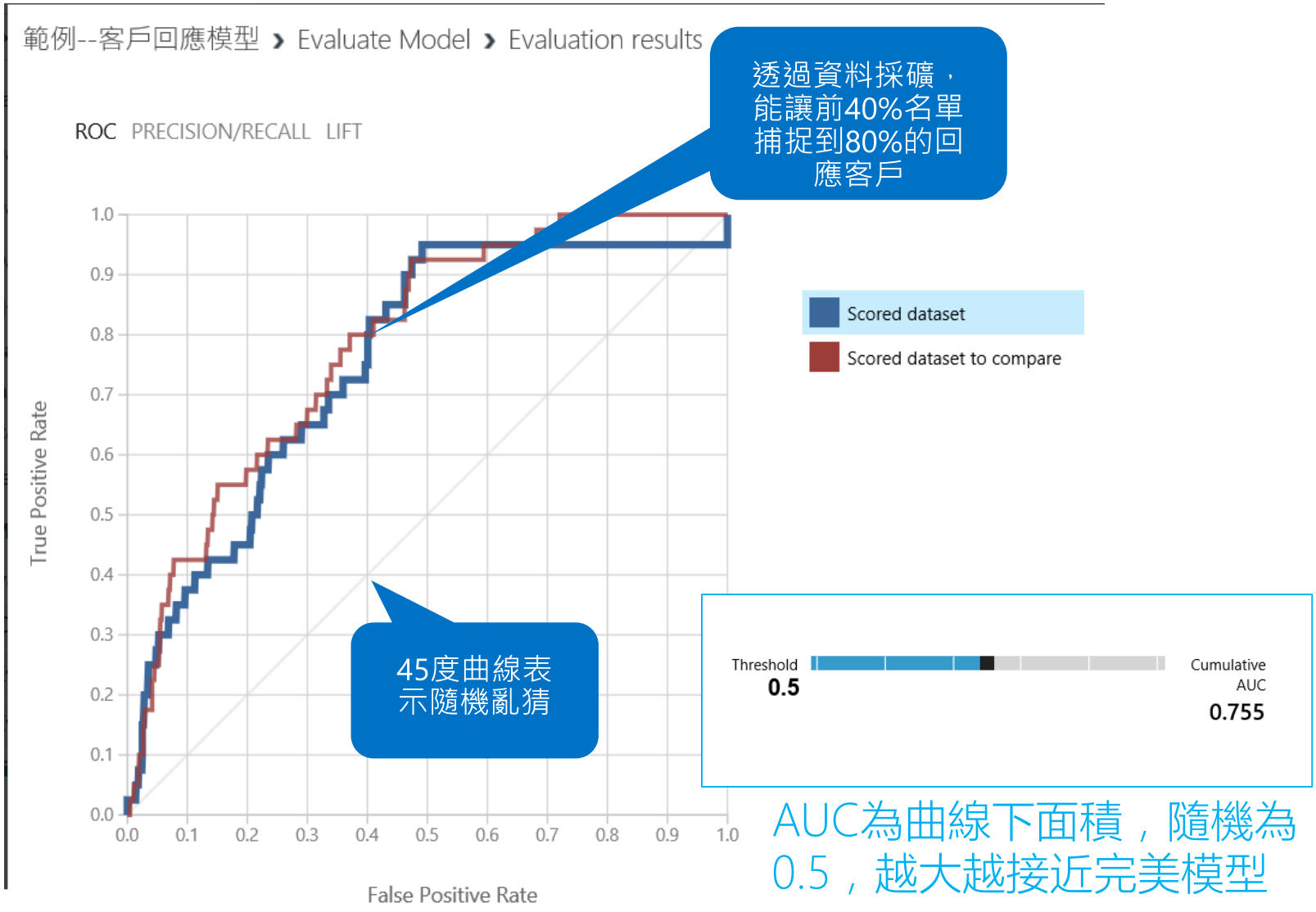
神經網路結構


X1
X2
X3
1
2
3
4
5
6
w14
w15
w24
w25
w34
w35
w46
w56
Y
輸入層
隱藏層
輸出層
Logistic(x)
= e(x) / (1 + e(x))
Weigted
Sum

產生預測結果



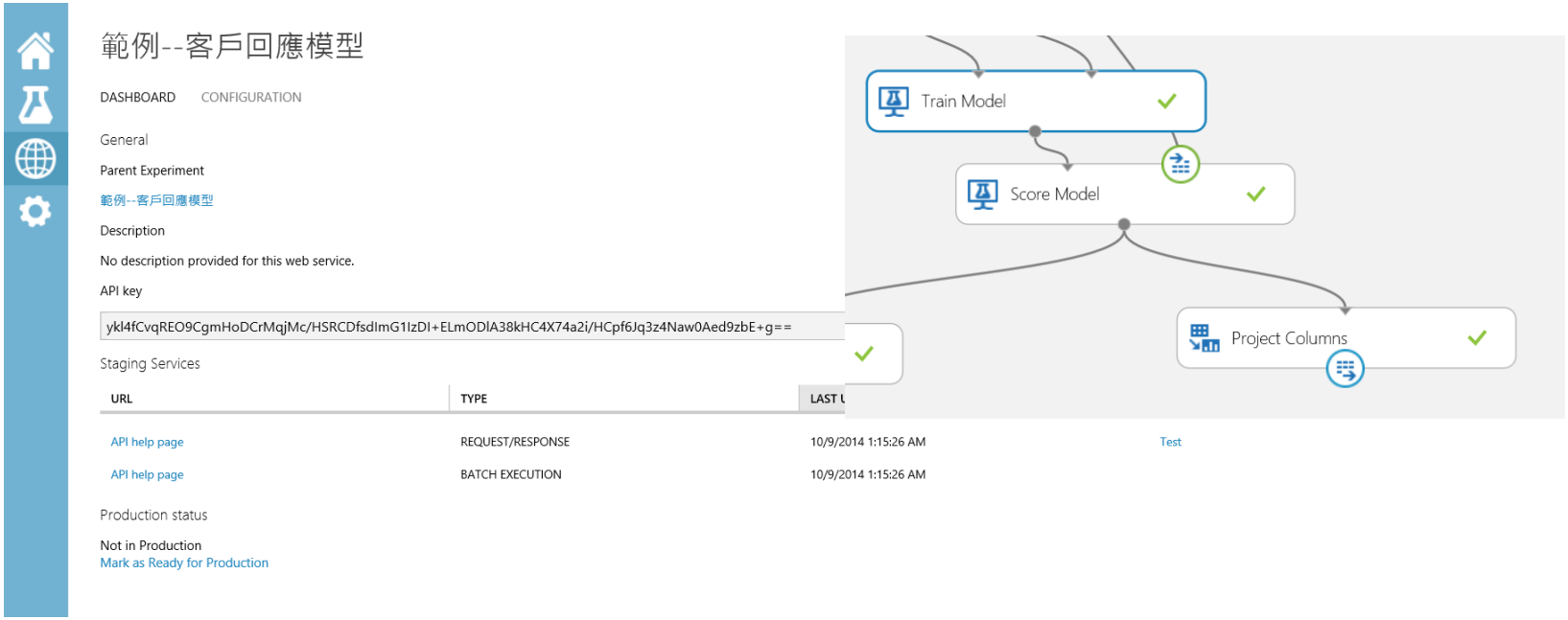
評估預測模型



AUC為曲線下面積，隨機為0.5，越大越接近完美模型

Prediction as a Service…

- 能將預測模型包裝為Web Services(單筆即時預測，多筆批次預測)
- 可透過C#, R以及Python等語言呼叫
- 將會是微軟物聯網策略中最重要的一環



Azure Machine Learning計費模式

客戶區隔模型

傾向模型與相似模型?

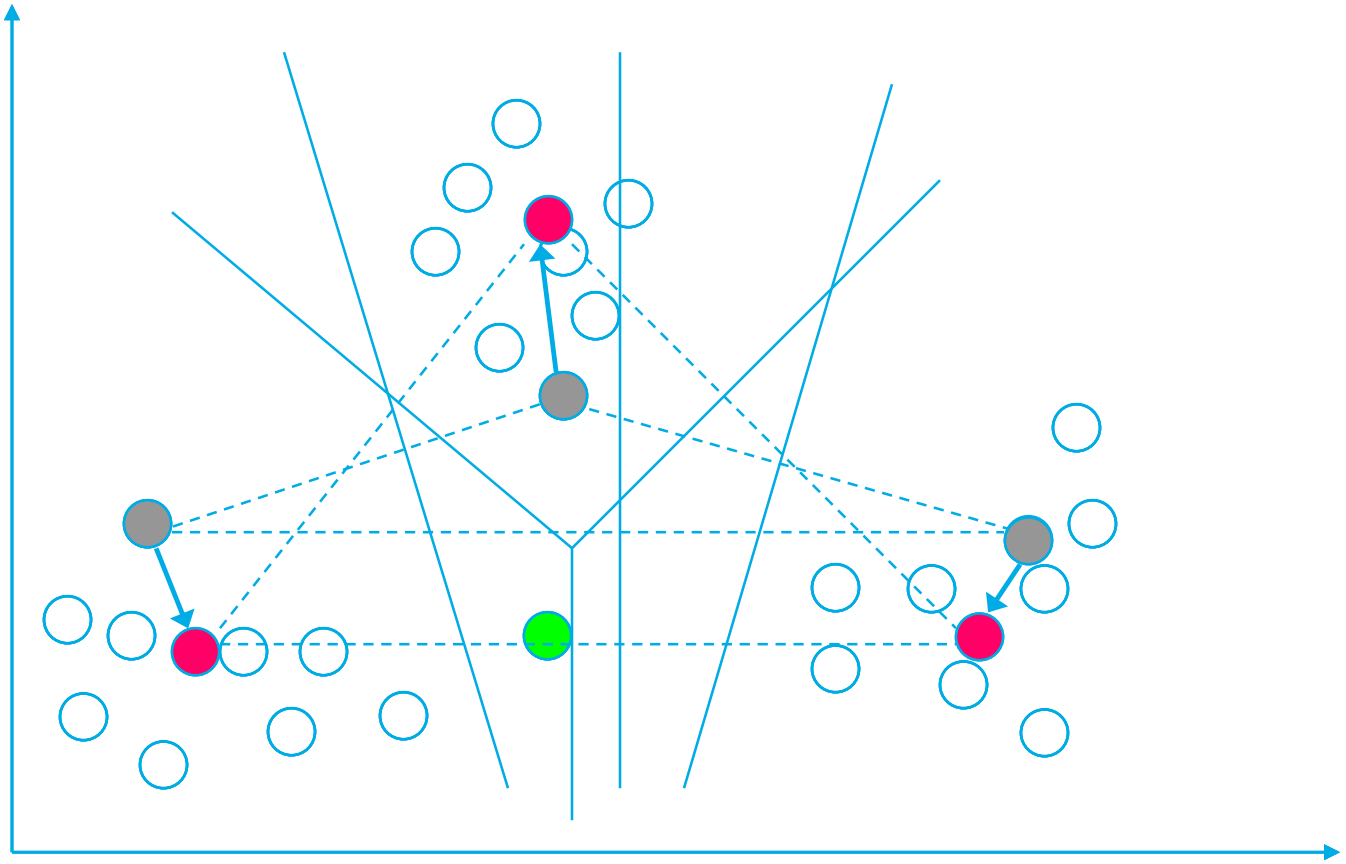
- 傾向模型(Propensity Model)
  - 輸入->輸出
- 相似模型(Look-like Model)
  - 沒有輸出變數
  - 找出相似結構...

Look-like Model

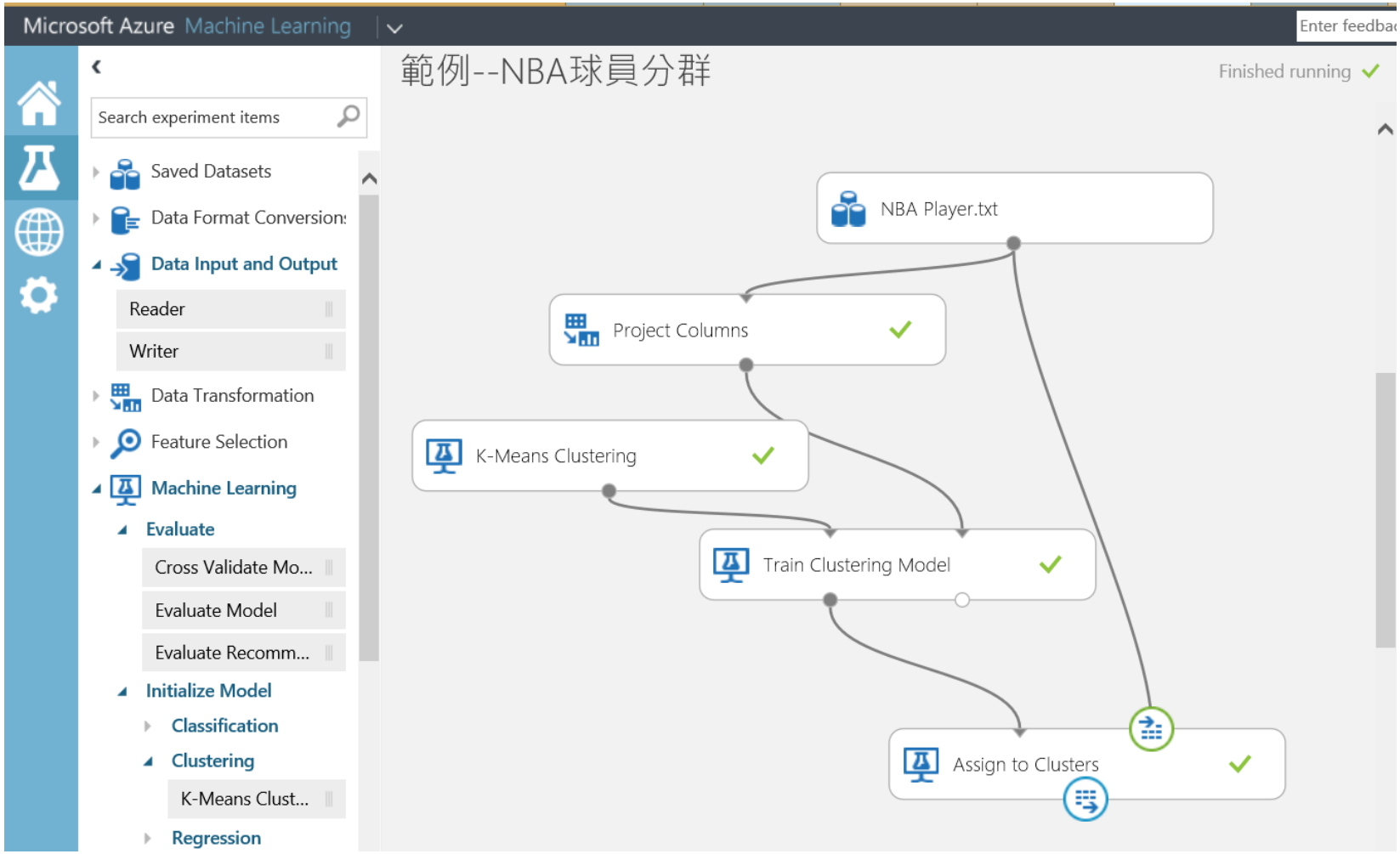
K-means演算法

- 最常使用的分群演算法
- K=分群數
- 透過隨機指派質量中心，透過不斷重新指派成員、計算新中心，來趨近於分群邊界

K-means



建立分群模型


Microsoft Azure Machine Learning
Enter feedba
Search experiment items
Saved Datasets
Data Format Conversions
Data Input and Output
Reader
Writer
Data Transformation
Feature Selection
Machine Learning
Evaluate
Cross Validate Mo...
Evaluate Model
Evaluate Recomm...
Initialize Model
Classification
Clustering
K-Means Clust...
Regression
範例--NBA球員分群
Finished running
NBA Player.txt
Project Columns
K-Means Clustering
Train Clustering Model
Assign to Clusters

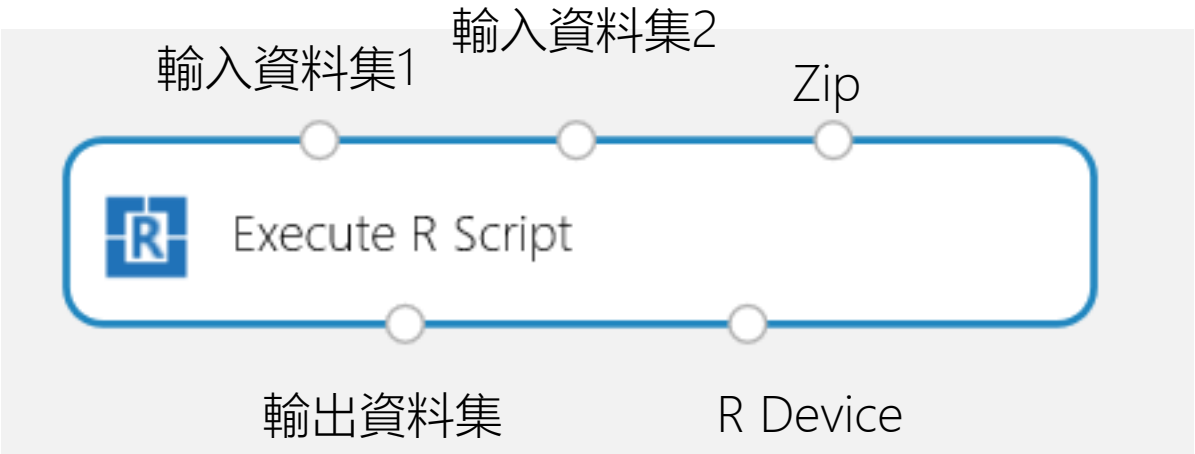
整合R語言

R語言

- R語言是由紐西蘭奧克蘭大學的Ross Ihaka和Robert Gentleman開發的開源統計語言
- 能夠進行統計、預測模型以及資料視覺化

http://www.r-project.org/

如何整合R語言



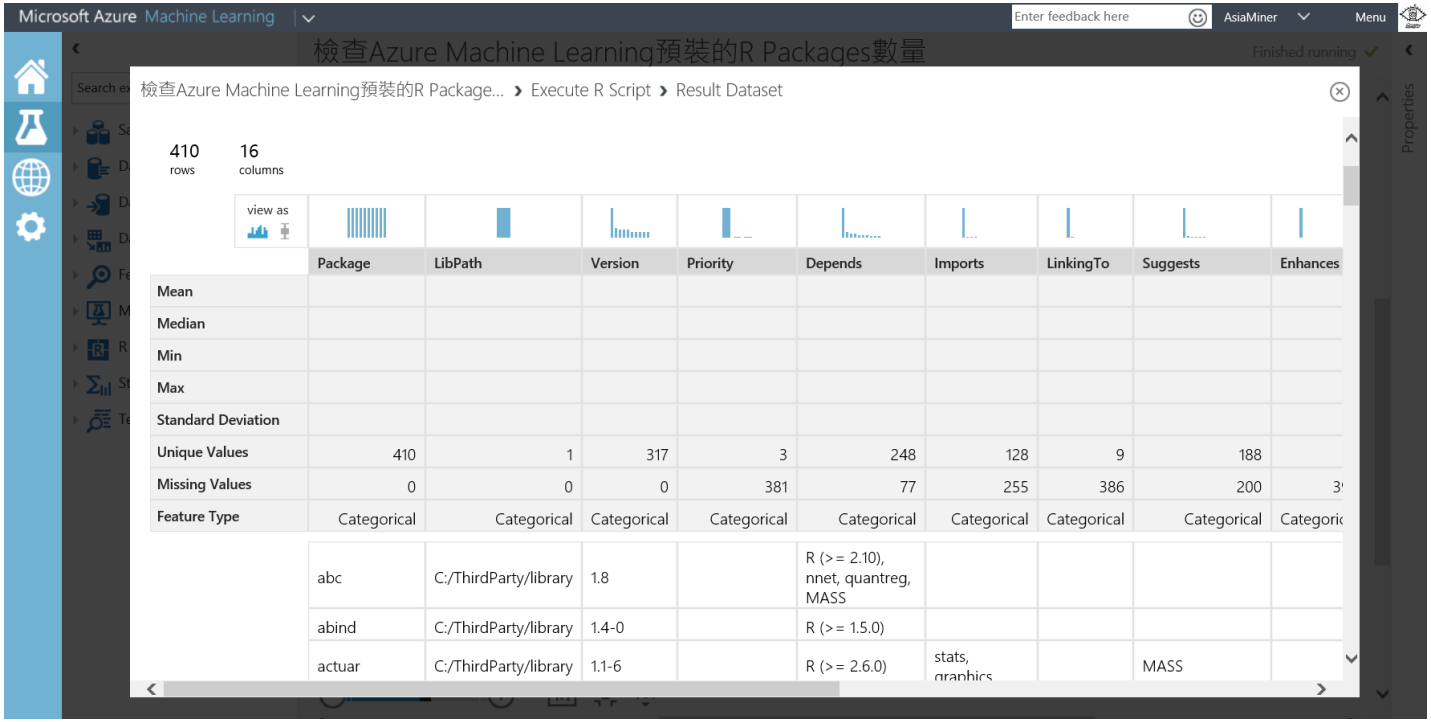

```
dataset1 <- maml.mapInputPort(1)
dataset2 <- maml.mapInputPort(2)

data.set = rbind(dataset1, dataset2);

maml.mapOutputPort("data.set");
```

輸入

輸出

檢視預裝的R Package

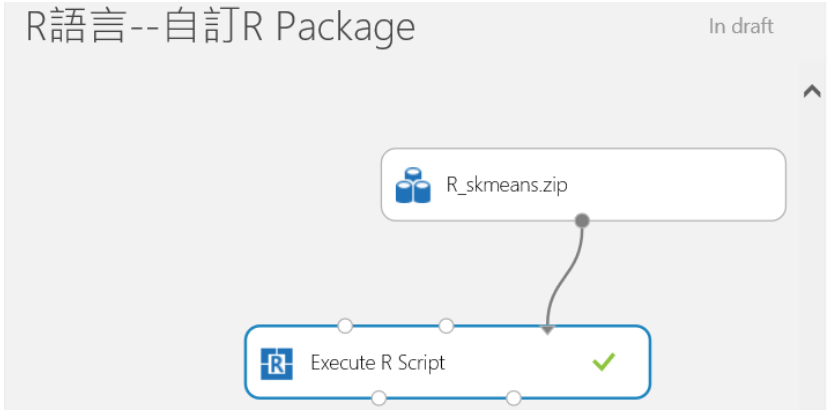
```
data.set<-data.frame(installed.packages())
maml.mapOutputPort("data.set");
```



	Package	LibPath	Version	Priority	Depends	Imports	LinkingTo	Suggests	Enhances
Mean									
Median									
Min									
Max									
Standard Deviation									
Unique Values	410	1	317	3	248	128	9	188	
Missing Values	0	0	0	381	77	255	386	200	3
Feature Type	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categorical	Categoric
	abc	C:/ThirdParty/library	1.8		R (>= 2.10), nnet, quantreg, MASS				
	abind	C:/ThirdParty/library	1.4-0		R (>= 1.5.0)				
	actuar	C:/ThirdParty/library	1.1-6		R (>= 2.6.0)	stats, graphics		MASS	

如何加入自訂R Packages

- 在電腦本機安裝想要的R Package
- 將R Package原始的Zip檔壓縮為Zip
- 上傳至Azure Machine Learning



Properties

Execute R Script

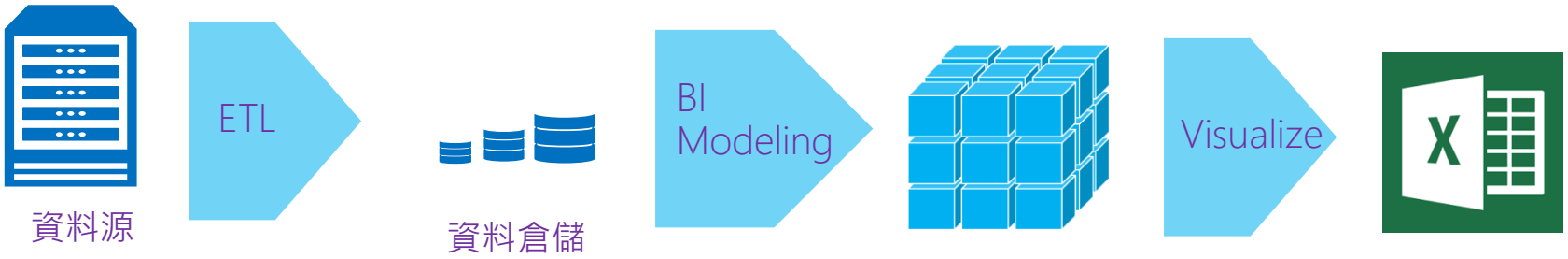
R Script

```
1 install.packages("src/slam_0.1-32.zip", lib = ".", repos = NULL, verbose = TRUE)
2 install.packages("src/clue_0.3-48.zip", lib = ".", repos = NULL, verbose = TRUE)
3 install.packages("src/skmeans_0.2-6.zip", lib = ".", repos = NULL, verbose = TRUE)
4 library(skmeans, lib.loc=".", verbose=TRUE)
5 success <- library("skmeans", lib.loc = ".", logical.return = TRUE, verbose = TRUE)
6 print(success)
```

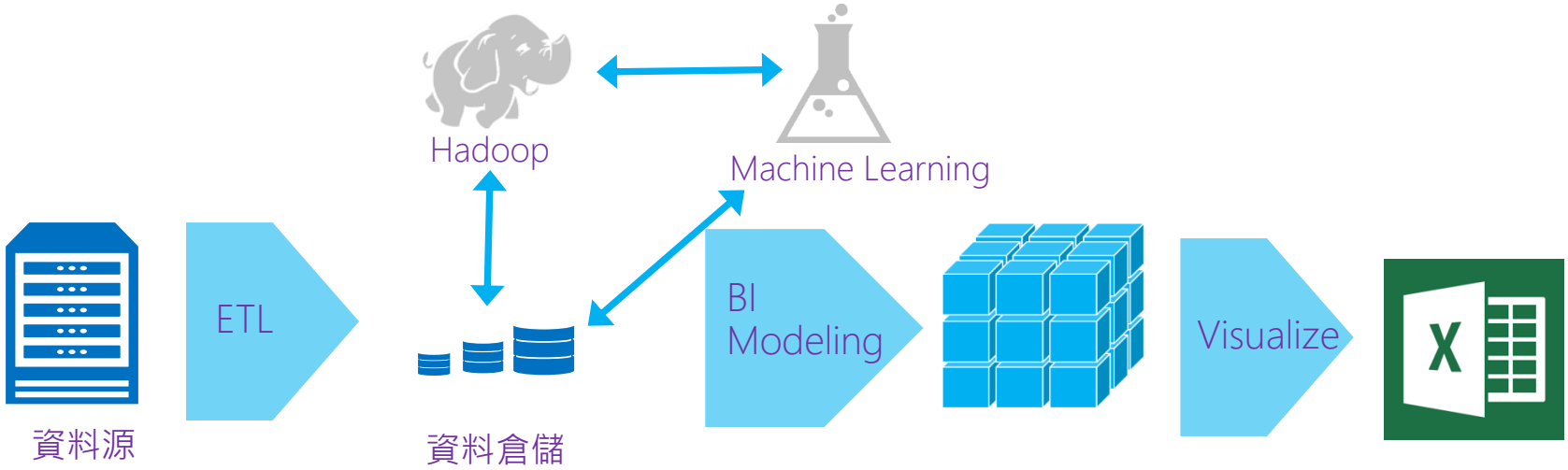
Random Seed

整合Power BI for Office 365

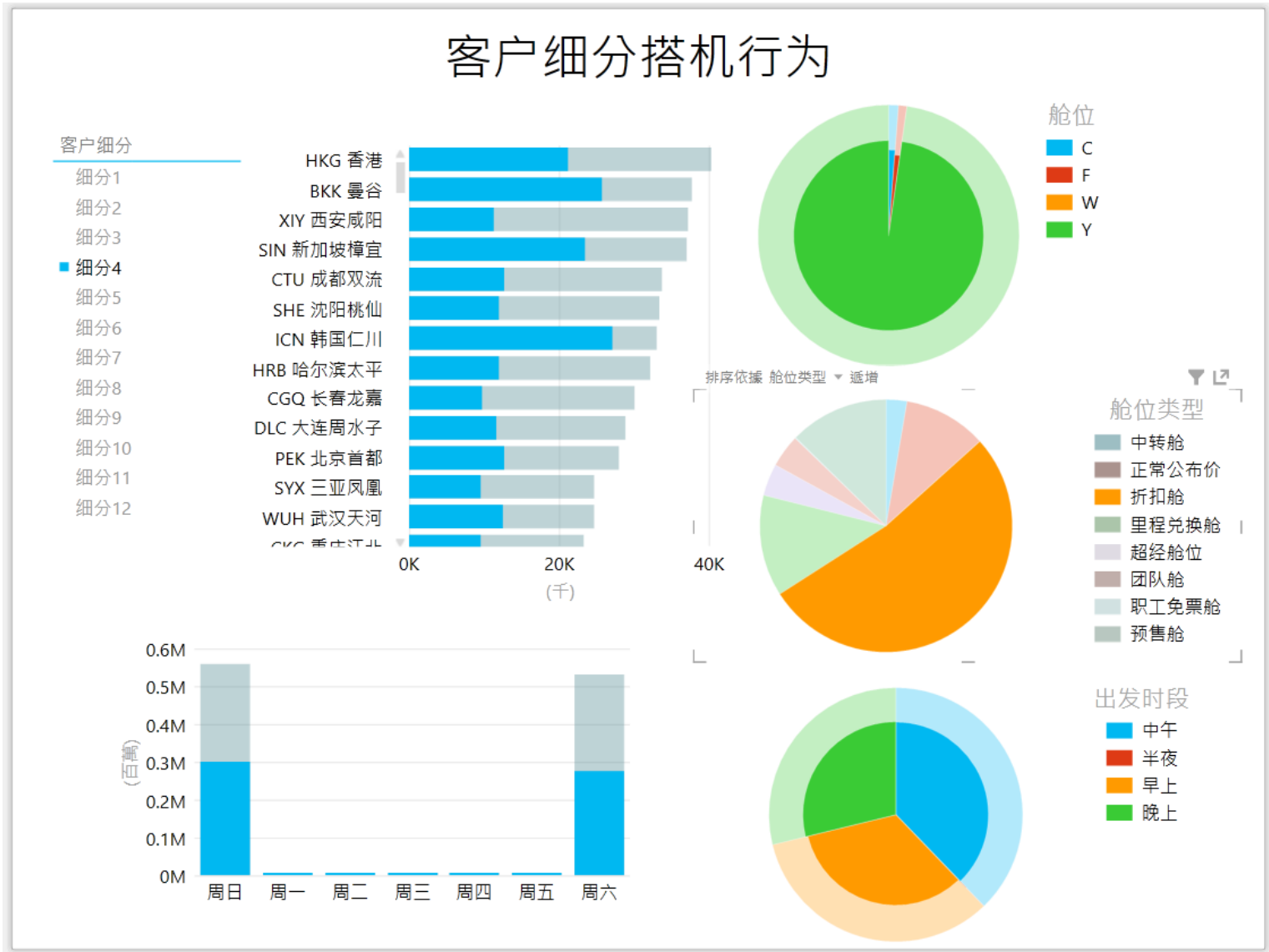
Traditional BI



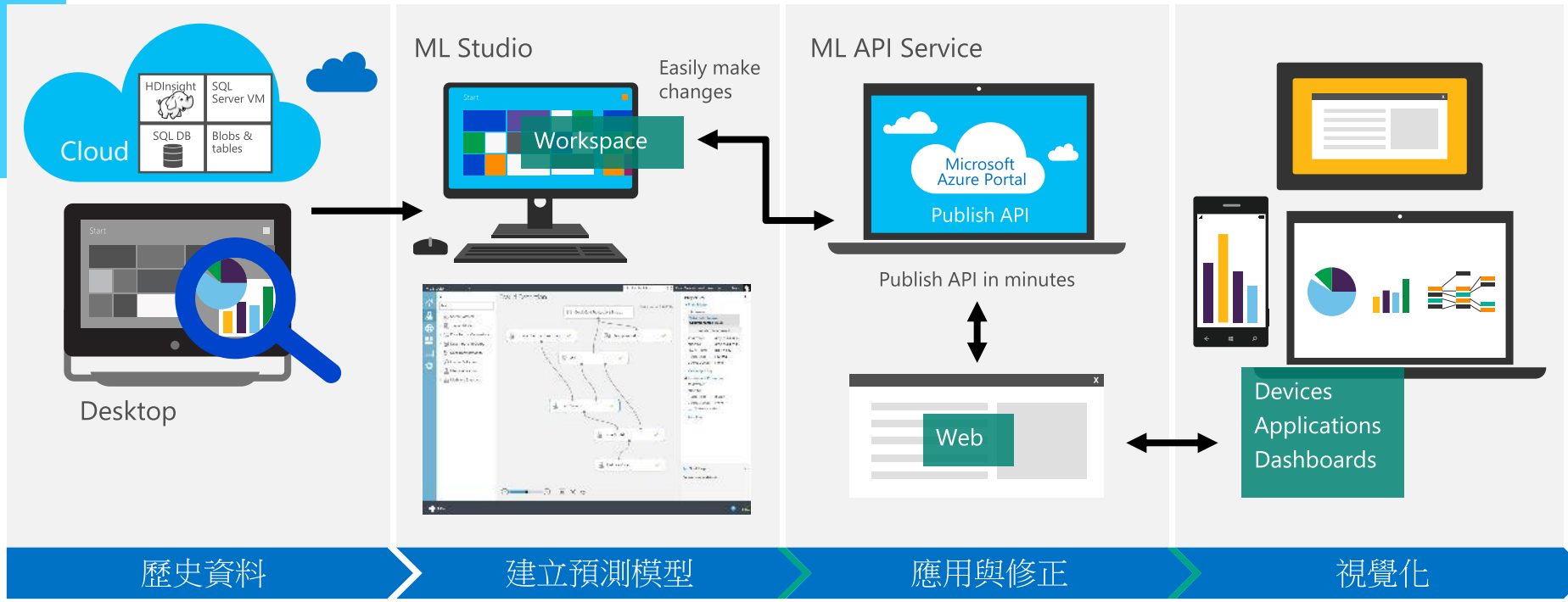
Modern Predictive BI



整合Power BI


客户细分搭机行为
客户细分
细分1
细分2
细分3
细分4
细分5
细分6
细分7
细分8
细分9
细分10
细分11
细分12
HKG 香港
BKK 曼谷
XIY 西安咸阳
SIN 新加坡樟宜
CTU 成都双流
SHE 沈阳桃仙
ICN 韩国仁川
HRB 哈尔滨太平
CGQ 长春龙嘉
DLC 大连周水子
PEK 北京首都
SYX 三亚凤凰
WUH 武汉天河
0K
20K
(千)
舱位
C
F
W
Y
排序依据 舱位类型
遞增
40K
舱位类型
中转舱
正常公布价
折扣舱
里程兑换舱
超经舱位
团队舱
职工免票舱
预售舱
0.6M
0.5M
0.4M
0.3M
0.2M
0.1M
0M
(百萬)
周日
周一
周二
周三
周四
周五
周六
出发时段
中午
半夜
早上
晚上

完整的數據分析生命週期



Q&A