

MACHINE LEARNING WITH R

*Machine Learning
Training and Testing
Pipelines*

胡中興 博士



行雲流水軟體(*AppCloudom Software*) 執行長
*bigDataSpark*大數據論壇 召集人

財團法人國家實驗研究院 台灣海洋科技研究中心
Taiwan Ocean Research Institute (TORI) 5/10 ~ 5/12/2017

胡 中 興 博士

President & CEO, AppCloudom Software

美國加州大學洛杉磯分校 (UCLA) 航太工程博士

- **Android/iOS App 開發, Big Data Analytics, AI**
- **C/C++, Java, R, SA&D, Networking, Security**

- 現職
 - 行雲流水軟體開發股份有限公司 董事長暨執行長
 - bigDataSpark 大數據論壇 召集人
- 外聘講師
 - 精誠資訊/高雄恆逸資訊教育訓練中心
 - 勞動部 高屏澎東分署 職訓中心
- 國立高雄第一科技大學 資管系 兼任助理教授
- 國立東華大學 東部區域運輸發展研究中心 兼任研究員
- 2005~2012 專任助理教授
 - 美和科技大學 資訊科技系



AppCloudom Software Dev. Corp.

行雲流水軟體開發股份有限公司



OUTLINE

1. INTRODUCTION
2. AI Learning Agents
3. Why Machine Learning (ML)?
4. ML Pipeline in Big Data Analytics
5. Machine Learning with R
6. Concluding Remarks

1. INTRODUCTION



AI & “Big Data”

- 德國 工業4.0 相關技術領域整合：
 1. 網際網路 (Internet)
 2. 行動網路
 3. 雲端運算
 4. 物聯網 (Internet of Things)
 5. 大數據分析 (Big Data Analytics)
 6. 人工智慧 (Artificial Intelligence)
 7. 智慧型機器人 (Intelligent Robots)
 8. 人機界面 & 系統與系統介面



IT (資訊科技)

Rainer Strack (賴內 · 斯達克):
2030 年勞動力危機令人震驚
— 如何從現在開始解決！
(TED Talk)

....

AI & Big Data

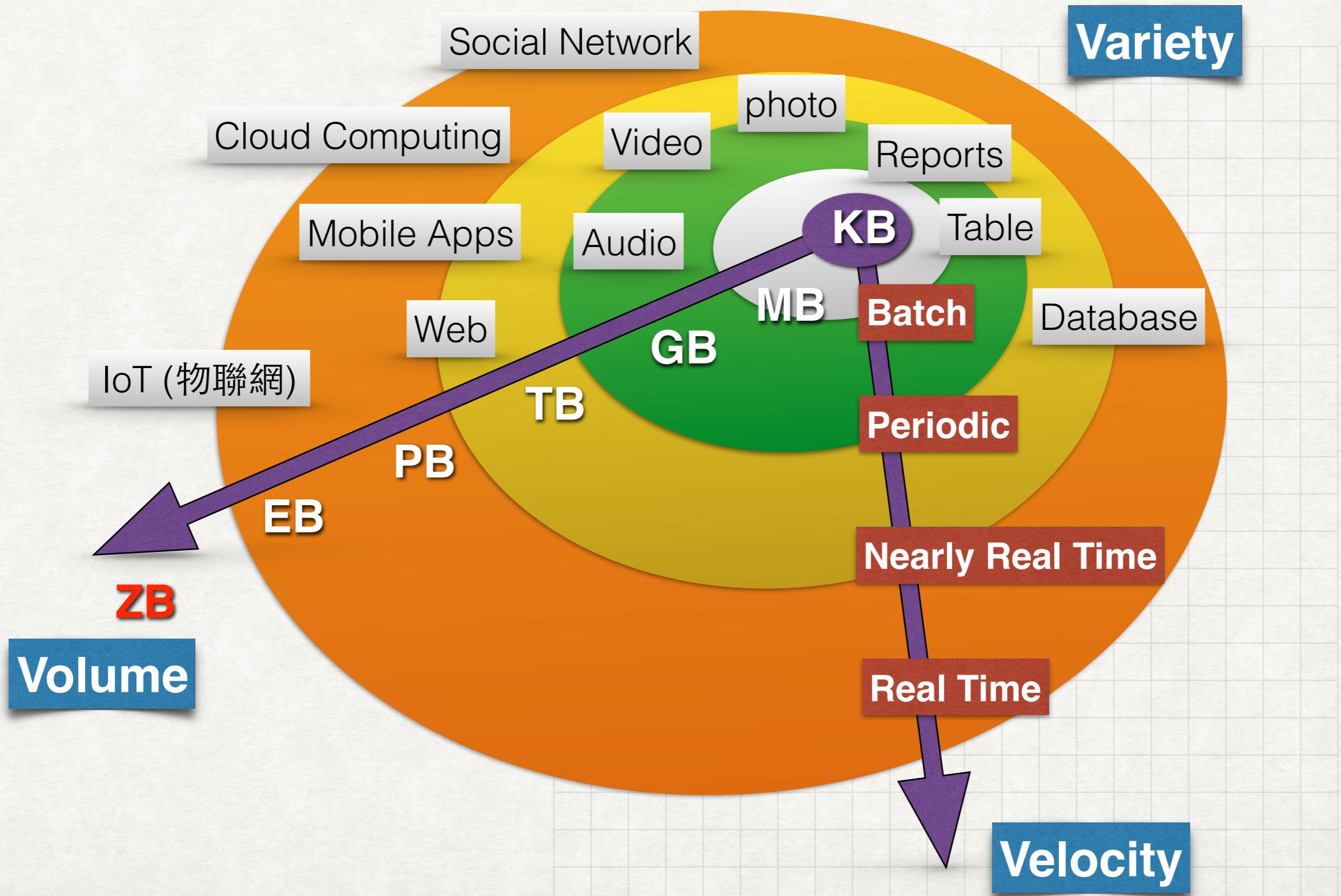
IPv6 for IoT Devices :

- 2^{128} devices = $2^8 \times (2^{10})^{12} \sim 256 \times (10^3)^{12}$
 $\sim 10^{38}$ devices  一百兆兆兆個裝置 !!
- 若每個 IoT 裝置：每一個月上傳 1GB (10^9 Bytes) 資料量
則一兆個裝置 (10^{12} devices)：每一個月將產生 1ZB (10^{21} Bytes) 資料量！

Q : 什麼時候“物聯網裝置”將會普遍使用呢？

A : Blockchains + IoT → Big Data Analytics

3V's Model of Big Data



*Case Studies in **Big Data Analytics***

- **Uncertainty** Problems
- **Edward Thorp**
 - “Beat the Dealer” (1964) - Black Jack
 - “Beat the Market” (1968) - Kelly Eqn.
- **“Money Ball” — Sabermetrics**
- **Netflix — *Movie Recommendation System***

工業4.0 – 大數據分析

大數據分析案例

- 節錄自

“大數據 **BIG DATA : A Revolution That Will Transform How We Live, Work, and Think”**

作者 麥爾荀伯格/ 庫基耶

(Viktor Mayer-Schonberger/ Kenneth Cukier)

譯者：林俊宏

出版社：天下文化

2013



大數據分析案例一

《案例一：2009 H1N1 流感大流行》

2009年，一種新的流感病毒 H1N1，結合了禽流感和豬流感病毒。短短幾個星期，全球的公共衛生機構都擔心即將爆發流感大流行。當時，還沒有能派上用場的疫苗，公共衛生當局唯一能努力的，就是減緩流感蔓延的速度。為了達到這項目的，必須先知道當前流行感染的範圍及程度。

就在 H1N1 躍上新聞頭條的幾星期前，網路巨擘谷歌(Google)旗下的幾位工程師，在著名的《自然》科學期刊發表了一篇重要的論文。該篇論文解釋了谷歌能如何「預測」美國在冬天即將爆發流感，甚至還能精準定位到是哪些州。

大數據分析案例二

《案例二：省錢方式的機票預訂系統》FARECAST

伊茲奧尼花了四十天，從某個旅遊網站取得超過一萬兩千筆票價資料，作為樣本，並建立一個預測模型，讓模擬的乘客都省下了大筆鈔票。這個模型並不懂「為何如此」(why)，只知道「正是如此」(what)。

換言之，模型完全不知道各種影響票價的因素，像是未售出的機位數、淡旺季、或是星期幾的機票較便宜之類；模型所做的預測，都是基於手中確實的資訊，也就是從其他航班所蒐集到的相關資料。

What is Data Scientist (資料科學家) ?

- 統計學家 (Statistician) vs. 資料科學家 (Data Scientist)
 - ❖ 統計學家通常利用三門數學學科的領域知識：
 1. 數據分析 (Data Analytics) — (SPSS, SAS, Excel...)
 2. 機率 (Probability)
 3. 統計 (Statistics)
 - ❖ 資料科學家需要的領域知識：
 - 數據分析、機率、統計、**程式設計(例：R, Python, Scala...)**、
平行運算、資料庫、機器學習 (Machine Learning) or
資料探勘 (Data Mining)
 - + 至少一門專業的領域知識 (製造業、銀行業、股票市場...)

2018年 五大資料經濟需求分佈

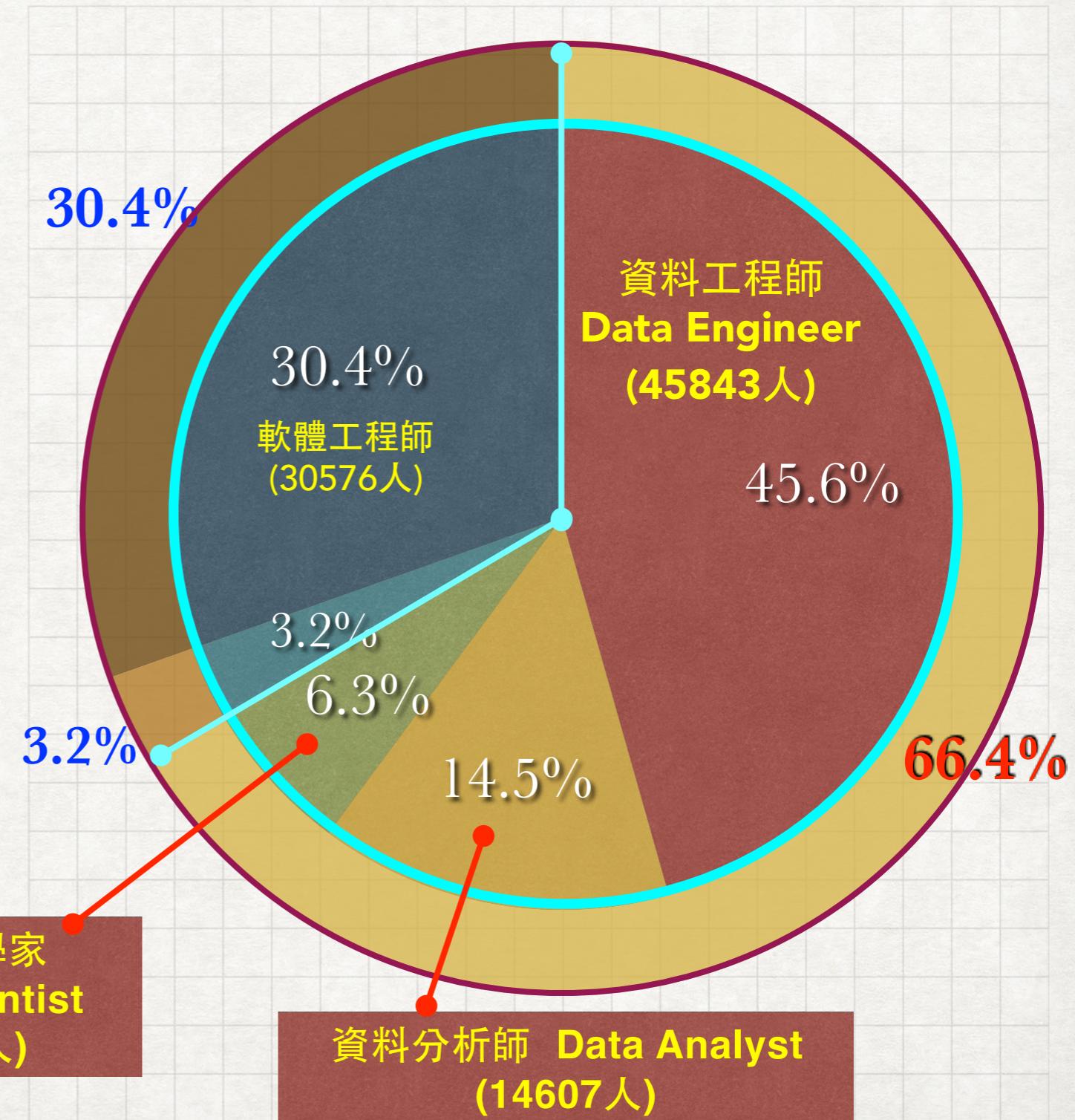
(依據 104 人力銀行 2016/1/2 資料經濟職務需求 調查結果 繪製)
(<http://www.ithome.com.tw/article/101983>)

調查結果顯示

1.	資料工程師	45843	(45.6%)
2.	軟體工程師	30576	(30.4%)
3.	資料分析師	14607	(14.5%)
4.	資料科學家	6296	(6.3%)
5.	領域專家	3201	(3.2%)

資料工程師 45,843 人
+ 資料分析師 14,607 人
+ 資料科學家 6,296 人

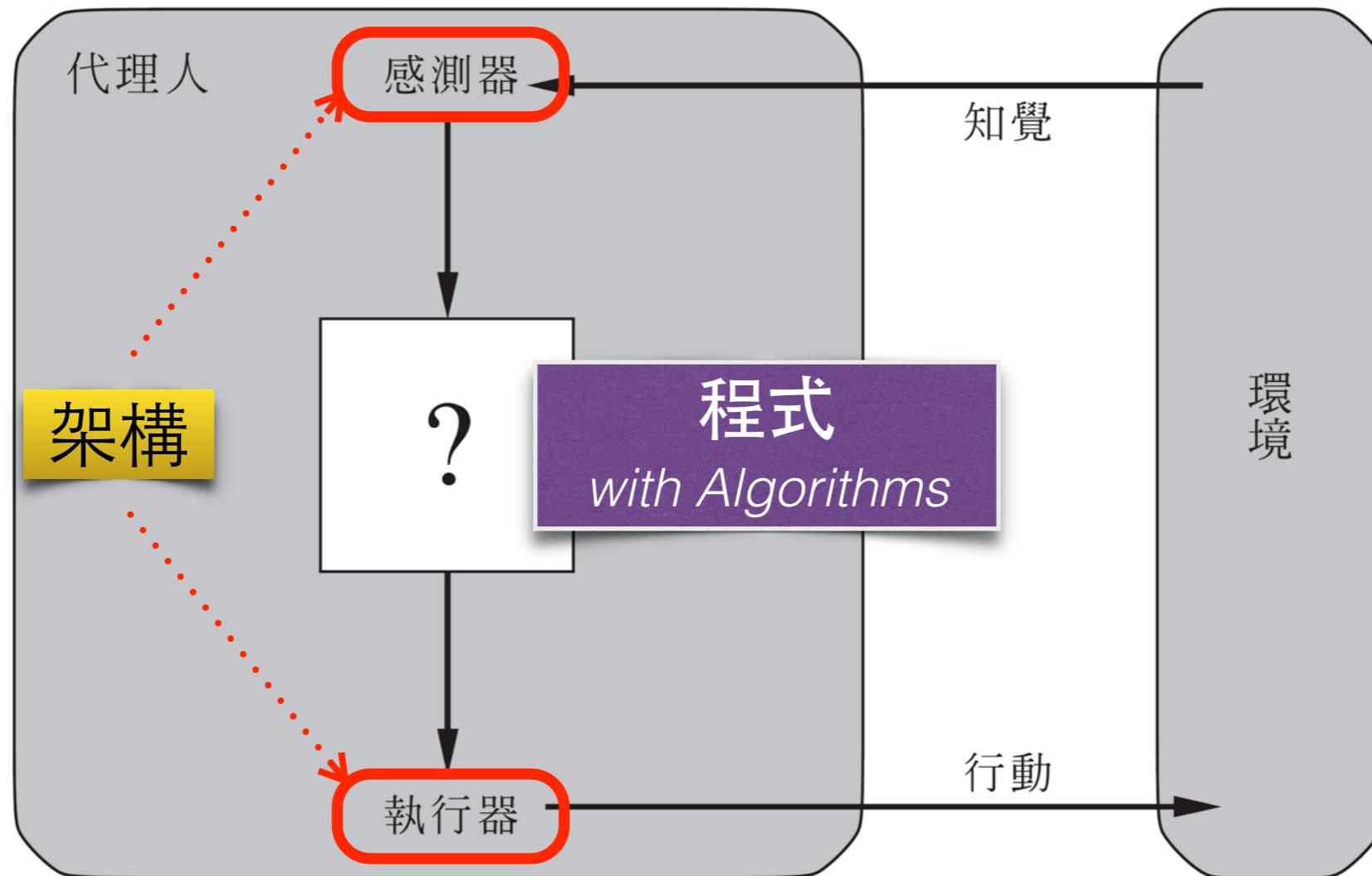
2018 總職缺需求 : 66,746 人 (66.4%)



2. AI Learning Agents

AI 代理人 (agent) 的結構

代理人 = 架構 (硬體) + 程式 (軟體)



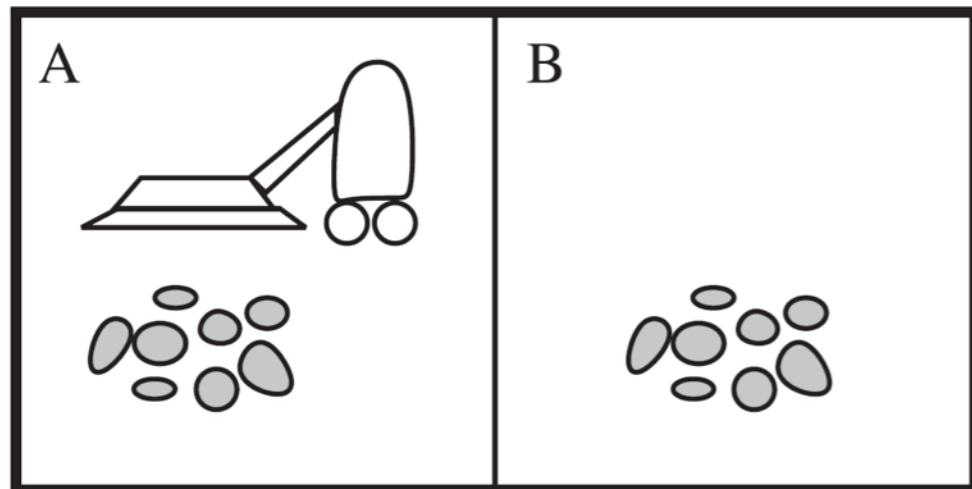
代理人透過感測器(sensors)和執行器(actuators)與環境進行交互作用

代理人程式

4 種基本的代理人程式 — 涵括了近乎所有智慧型系統的基礎原則：

- 簡單的反射型代理人 (simple reflex agents)
- 以模型為基礎的反射型代理人
- 基於目標的代理人
- 基於效用的代理人

Agent 1：簡單反射型代理人



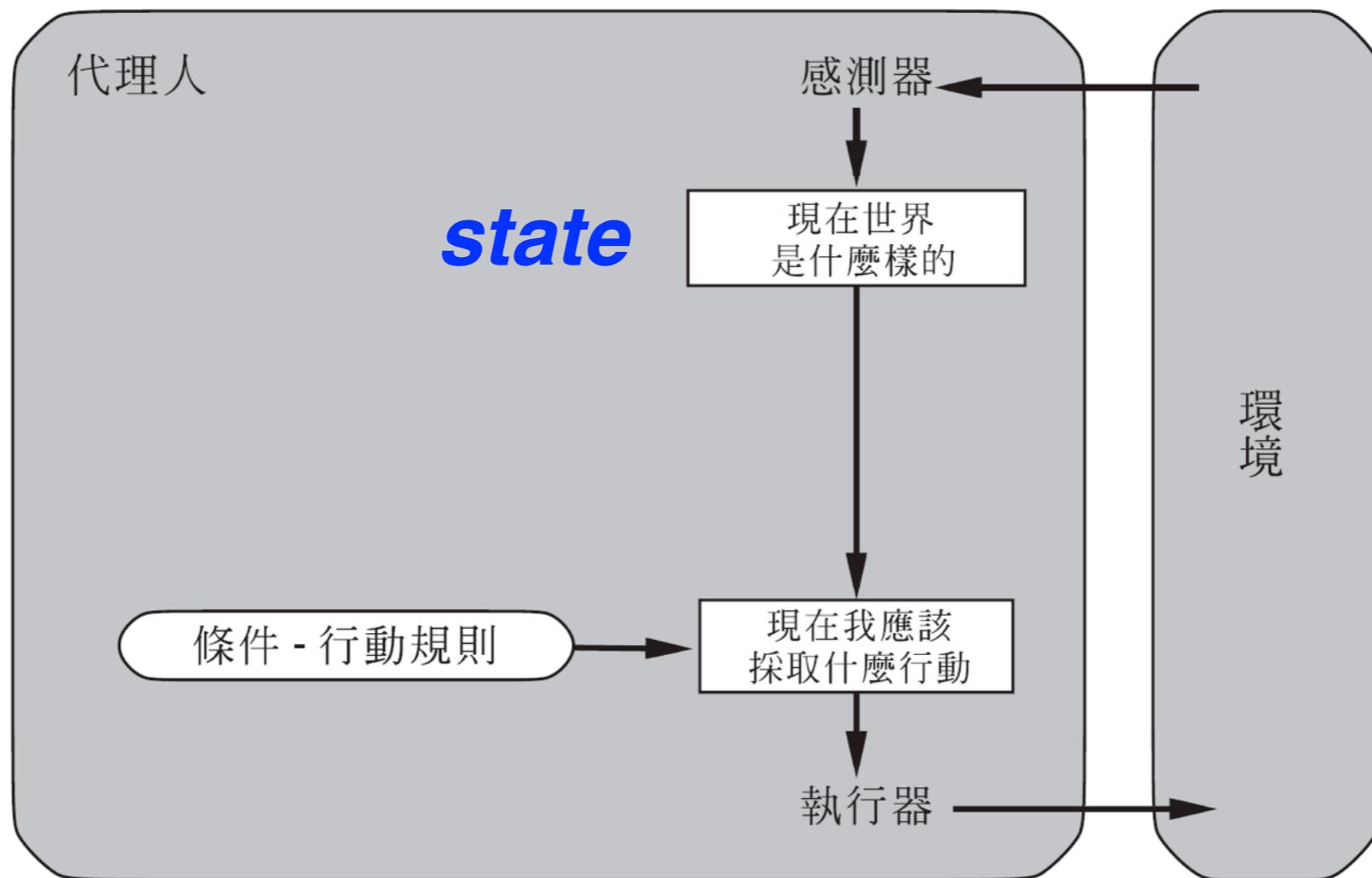
知覺序列	行動
[A, Clean]	Right
[A, Dirty]	Suck
[B, Clean]	Left
[B, Dirty]	Suck
[A, Clean], [A, Clean]	Right
[A, Clean], [A, Dirty]	Suck
:	:
[A, Clean], [A, Clean], [A, Clean]	Right
[A, Clean], [A, Clean], [A, Dirty]	Suck
:	:

兩狀態吸塵器環境中的簡單反射型代理人的代理人程式。（此程式實作右上表列的代理人函數）

```
function REFLEX-VACUUM-AGENT([location, status]) returns an action  
  
  if status = Dirty then return Suck  
  else if location = A then return Right  
  else if location = B then return Left
```

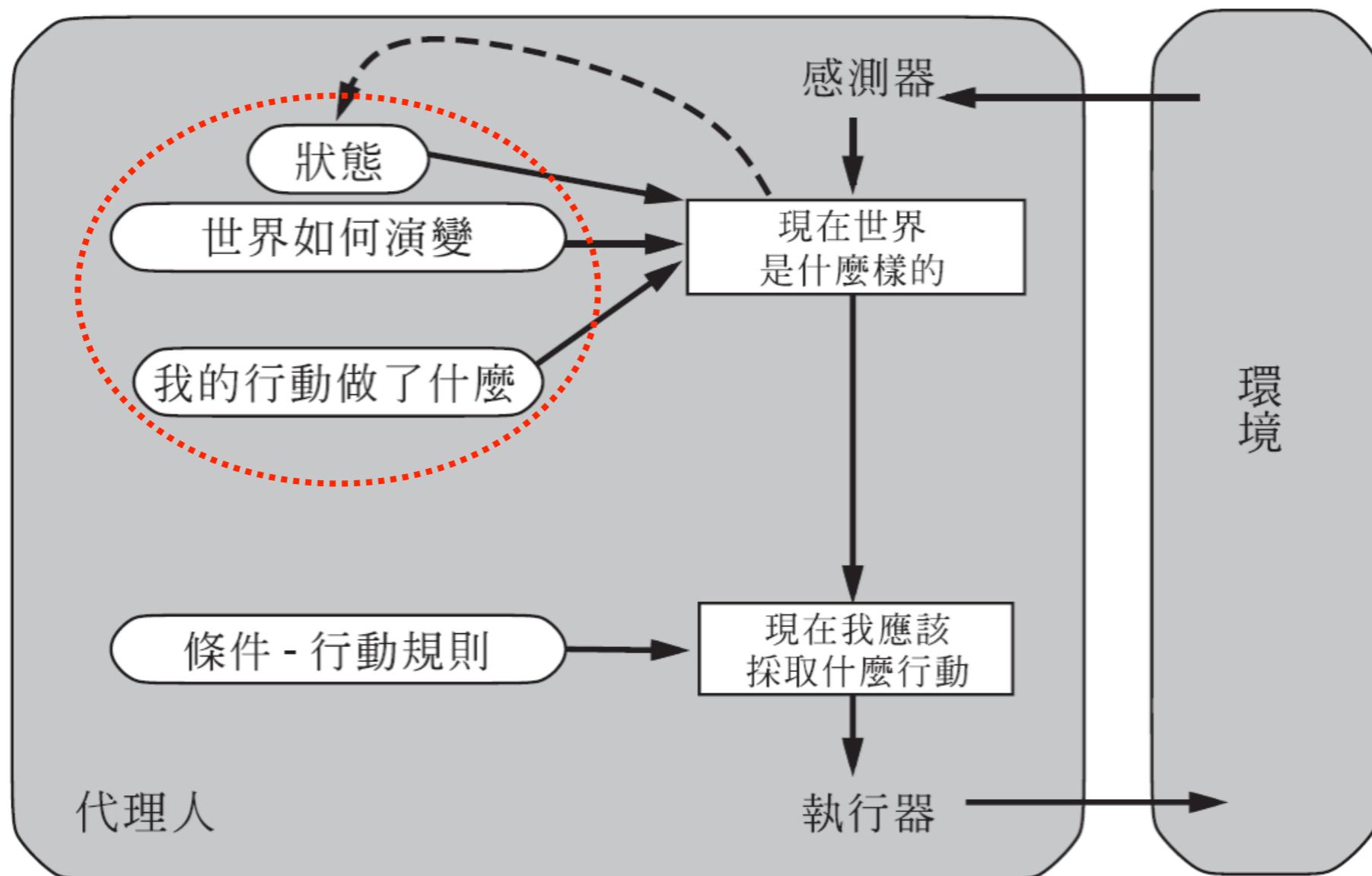
Agent 1：簡單反射型代理人 (cont'd)

簡單的反射型代理人示意圖



條件 - 行動規則：“若 … ， 則 … ” (例：若 前方的車輛在剎車， 則 開始剎車。)

Agent 2：基於模型的反射型代理人



Agent 2：簡單的反射型代理人 (cont'd)

基於模型的反射型代理人

- 它使用一個內部模型追蹤記錄世界的當前狀態。
- 然後，它採用與反射型代理人同樣的方式選擇行動。

function REFLEX-AGENT-WITH-STATE(*percept*) **returns** an action

static: *state*, a description of the current world state

rules, a set of condition-action rules

action, the most recent action, initially none

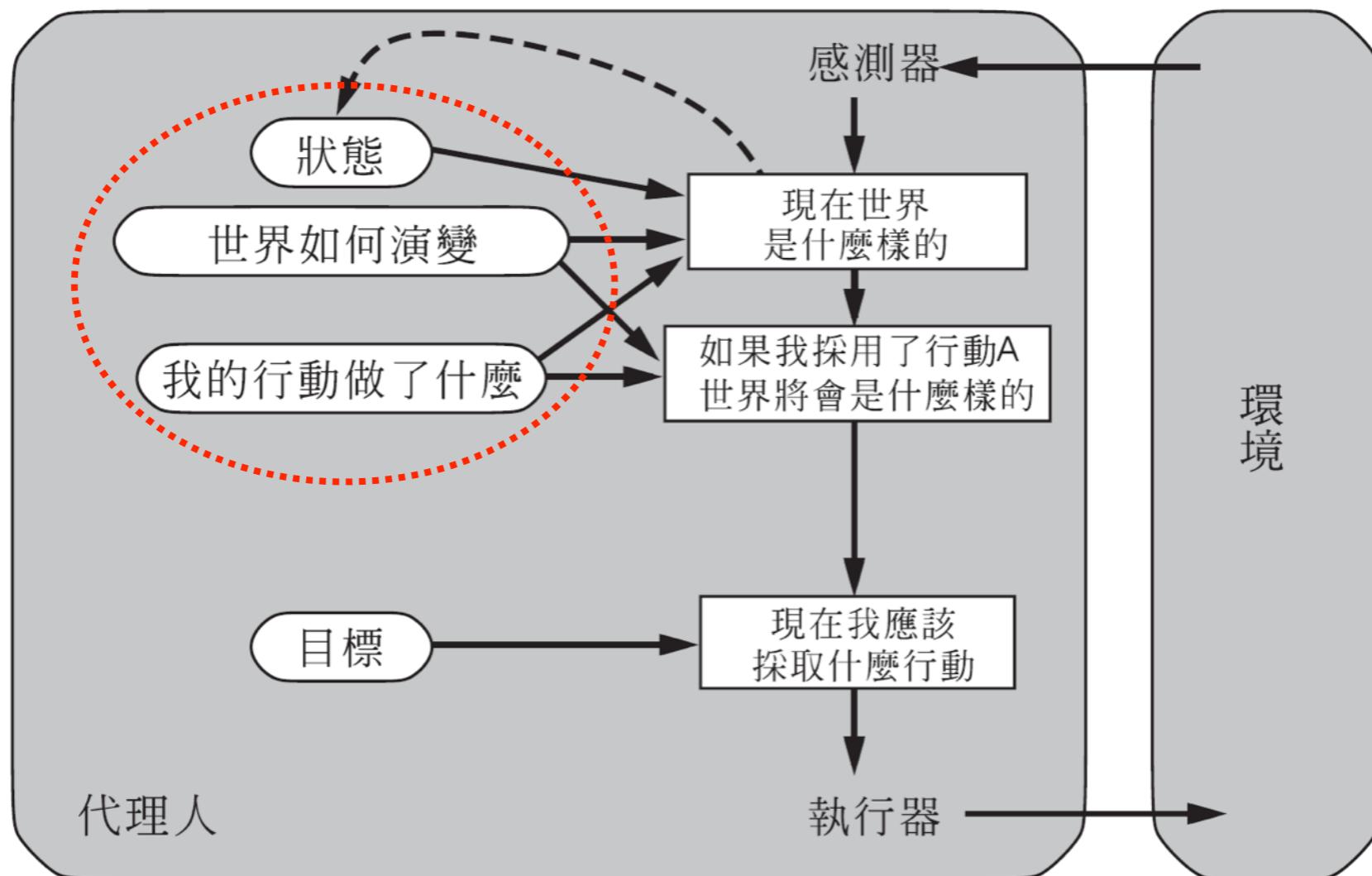
state \leftarrow UPDATE-STATE(*state*, *action*, *percept*)

rule \leftarrow RULE-MATCH(*state*, *rules*)

action \leftarrow RULE-ACTION[*rule*]

return *action*

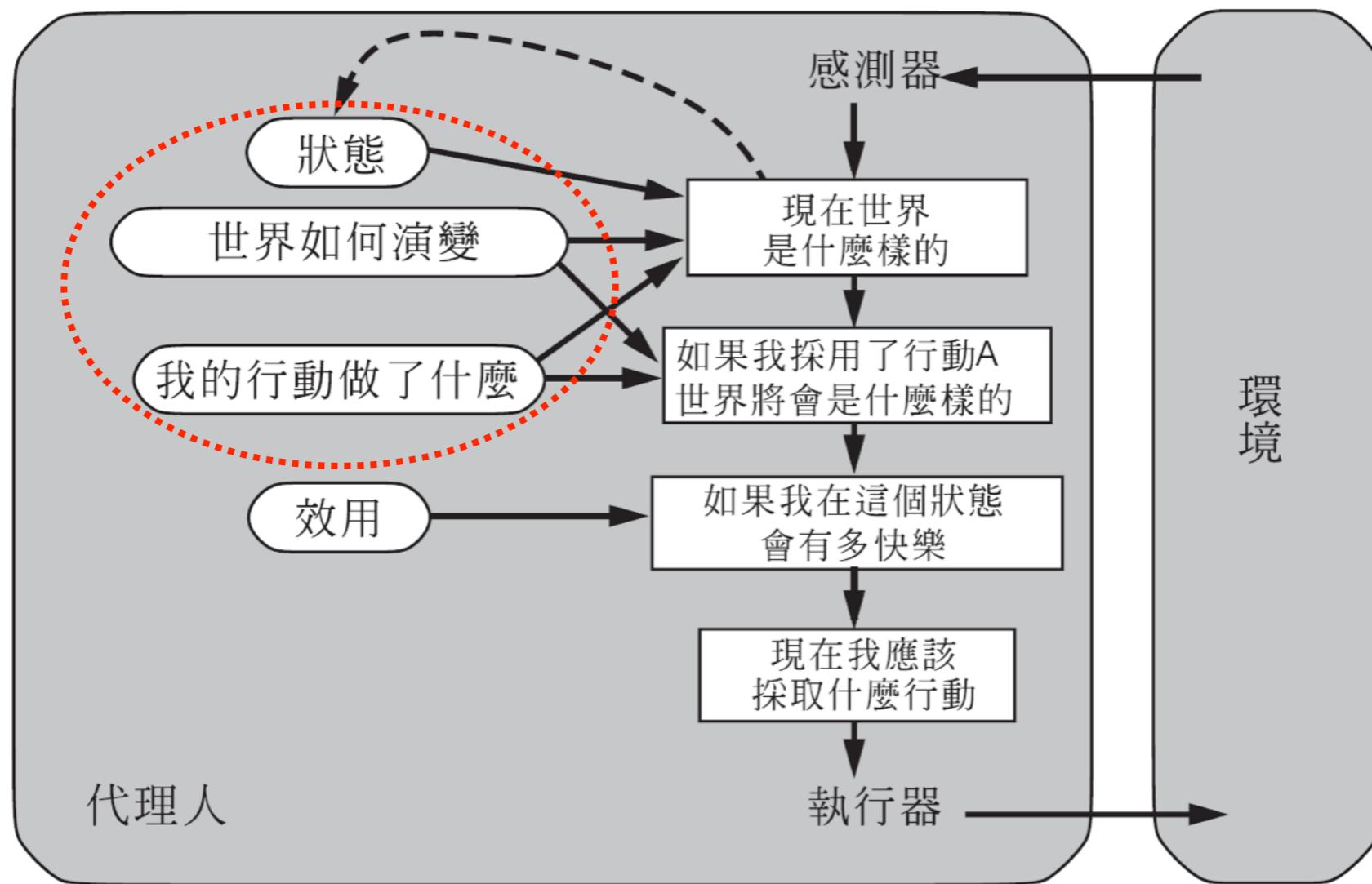
Agent 3：基於目標的代理人



基於模型和目標的反射型代理人 —

它既追蹤記錄世界的狀態，也記錄它要達到的一組目標，並選擇(最終)能導致其目標被達成的行動。

Agent 4：基於效用的代理人



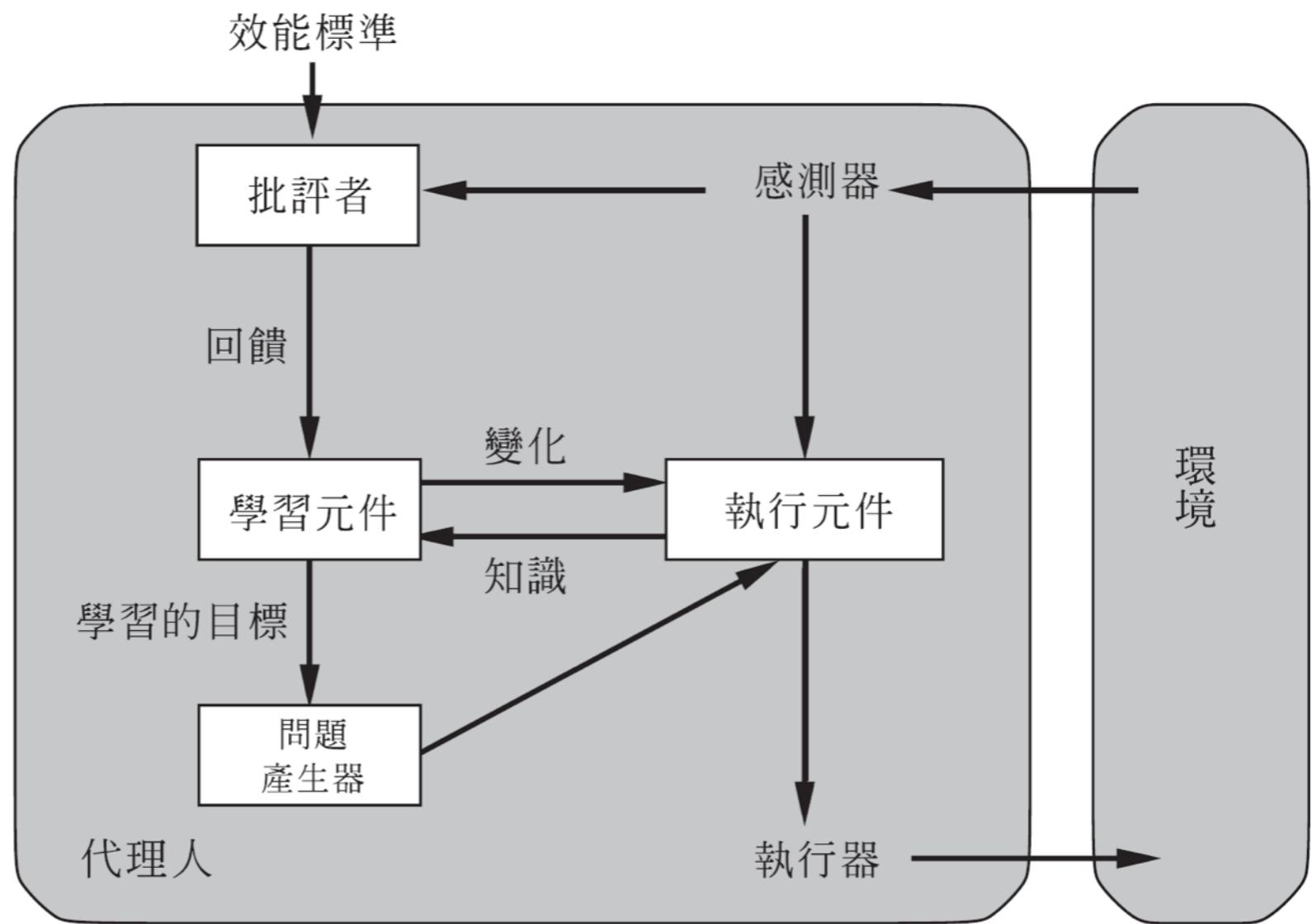
基於模型和效用的反射型代理人 —

- 它使用了一個世界的模型，以及一個評斷它對各個世界狀態的偏好程度的效用函數。
- 然後它選擇會導致最佳期望效用的行動。
- 最佳期望效用可藉由計算所有可能結果狀態的加權平均值得到，其權重為結果的機率。

一般的學習型代理人 (AI Learning Agent)

4 個概念上的元件：

- **學習元件** — 負責做出改進
- **執行元件** — 負責選擇外部動作
(相當於先前提過的代理人)
- **批評者** — 評價代理人做得如何
之後，**學習元件**利用來自批評者的
回饋來決定應該如何修改**執行元件**，使得在未來能夠做得更好
- **問題產生器** — 負責提出探索活
動，收集新資訊及經驗



Exploitation vs. Exploration

3. Why *Machine Learning* (ML)?

Q : Why Machine Learning (ML)?

Q1：為何我們要代理人學習？

Q2：如果代理人的設計可以被改進，為何設計者不開始時就根據可改進處來實作程式？

A：三個主要原因 —

- 設計者無法預知代理人可能發現自己所身處的所有可能情況。（例如，迷宮探險機器人）
- 設計者無法預知隨著時間的所有變化。（例如，預測股票市場的系統，需要學習適應調整）
- 設計者無法直接實作程式來解決的問題。（例如，機器對於人臉識別，需要間接透過學習演算法來達成）

機器學習 (Machine Learning)

三個類型的回饋(feedback) — 決定了學習的三個主要類型

- **無監督學習 (Unsupervised Learning)** : 代理人從輸入學習模式 (即使沒有提供明確回饋)
 - 群集 (clustering) 是最一般的無監督學習任務：偵測輸入例子中的潛在有用群集。
例如：無汽車教練的計程車代理人
- **受監督學習 (Supervised Learning)** : 代理人觀察一些輸入輸出對的例子，並學習由輸入對應到輸出的函數
例如：有汽車教練的計程車代理人
- **強化學習 (Reinforcement Learning)** : 代理人從一連串的強化 — 回報(reward)或懲罰(penalty) — 來學習。
 - 這取決於代理人判斷，強化之前的哪一個行動是造成強化的原因。



*R*hadoop = RHadoop

R + Hadoop = Big Data Analytics



**Statistics
+
Machine Learning
+
Data Visualization**

**HDFS / HBase
+
MapReduce**

for Big Data

for analytics

**Resolving the practical problems
(Website-popularity monitoring,
Stock-Market Prediction,
Business-Model Revealing...)**

**Spark
SQL**

**Spark
Streaming**

**MLlib
(Machine
Learning)**

**GraphX
(Graph)**

SparkR

Apache Spark

Apache Hadoop

Apache Spark Ecosystem

4. Machine Learning Pipeline in Big Data Analytics

資料前置處理
(Data Pre-processing)
☞ 需要“領域知識”

原始資料
Raw Data

傳統資料分析 (R)
vs.
Big Data 資料分析 (SparkR)

1

2

資料視覺化
(Data Visualization)

特徵向量擷取
(Feature Vector's
Determination)

資料分析方法：
• 機率模型
• 統計模型
• 資料探勘 (Data Mining)

傳統資料分析流程 : R

3

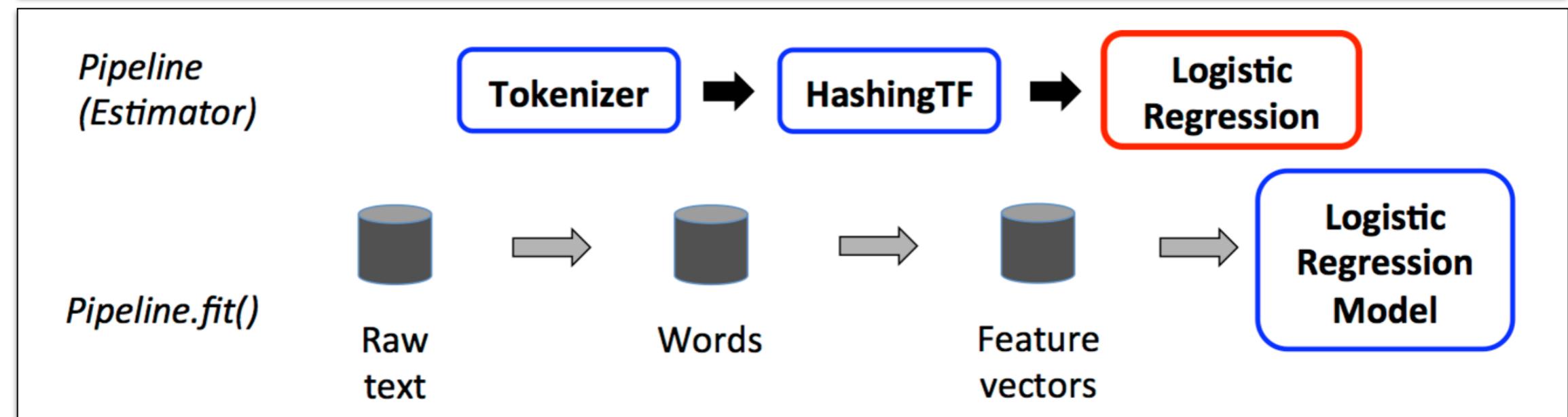
機器學習訓練與測試流程

(Machine Learning Pipeline for Training & Testing)

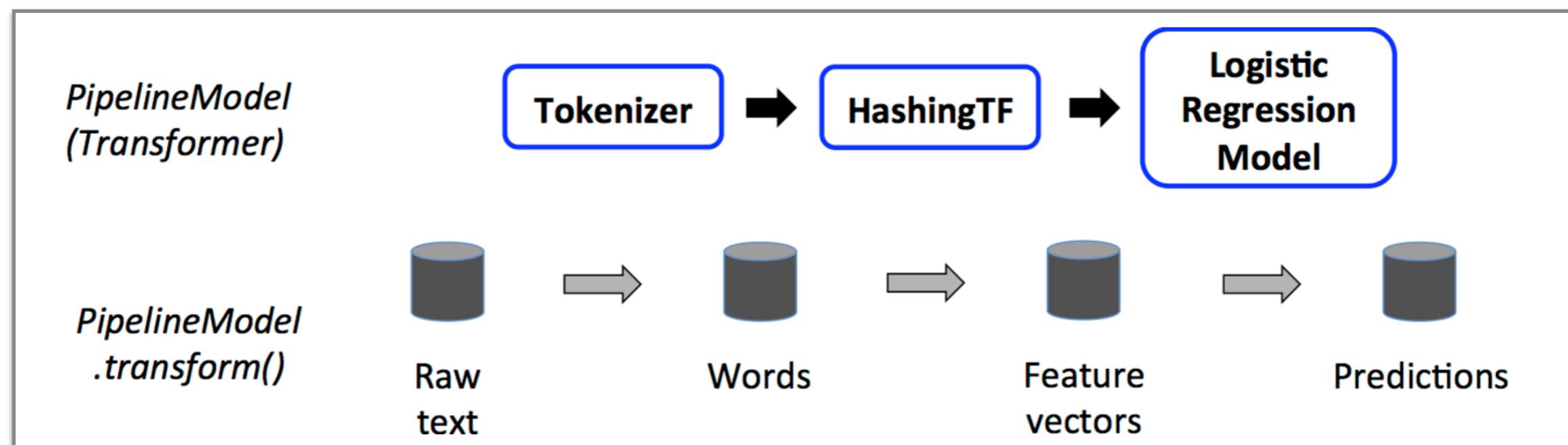
Big Data 資料分析流程 : SparkR

ML Pipeline on Spark (<http://spark.apache.org/docs/latest/ml-pipeline.html>)

The training time usage of a Pipeline — A Pipeline is an Estimator.



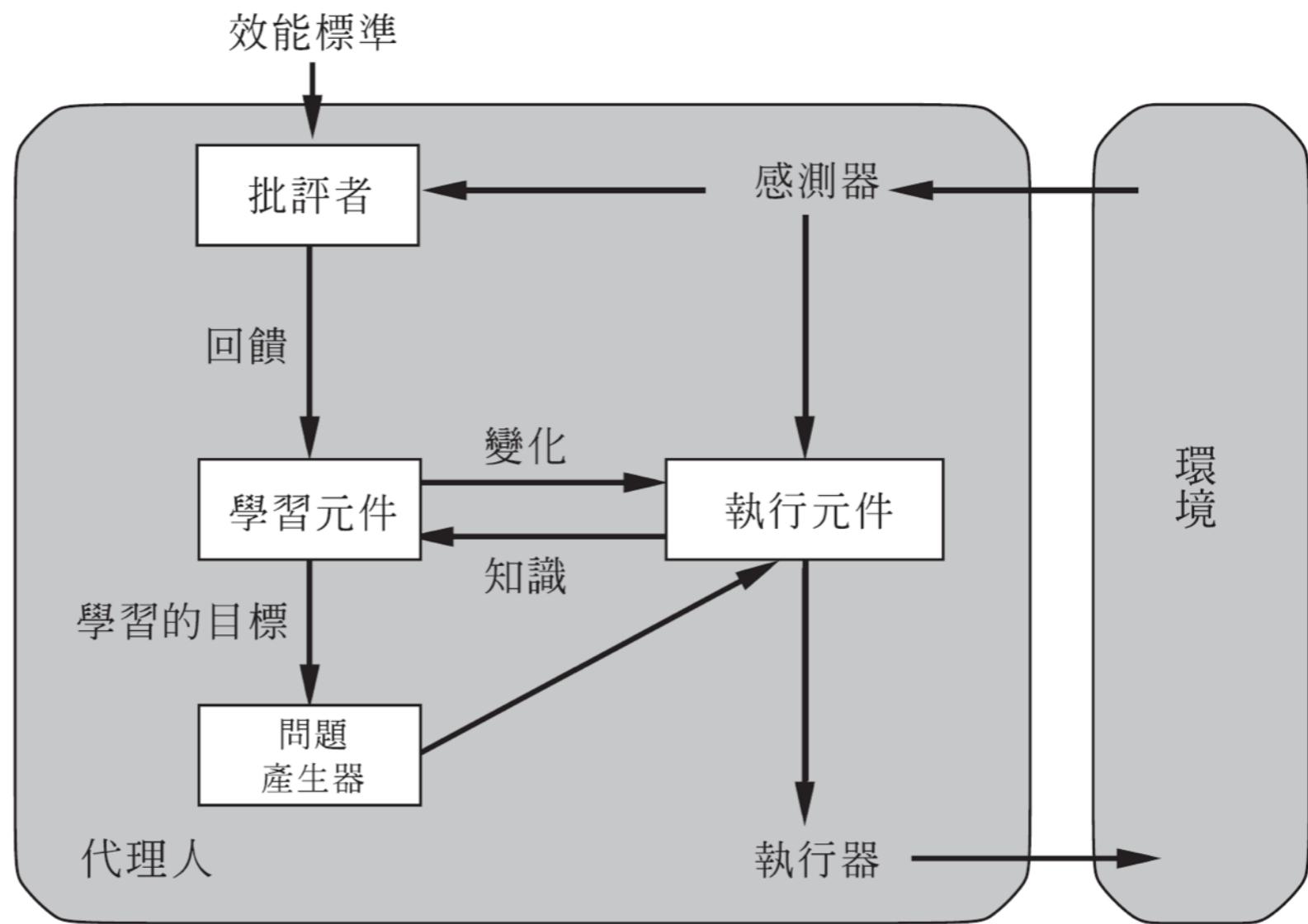
A PipelineModel is used at test time. — A PipelineModel is a Transformer.



導入 AI Learning Agent 概念 — 完整的機器學習模型

4 個概念上的元件：

- **學習元件** — 負責做出改進
- **執行元件** — 負責選擇外部動作
(相當於先前提過的代理人)
- **批評者** — 評價代理人做得如何
之後，**學習元件**利用來自批評者的
回饋來決定應該如何修改**執行元件**，使得在未來能夠做得更好
- **問題產生器** — 負責提出探索活
動，收集新資訊及經驗



Exploitation + Exploration

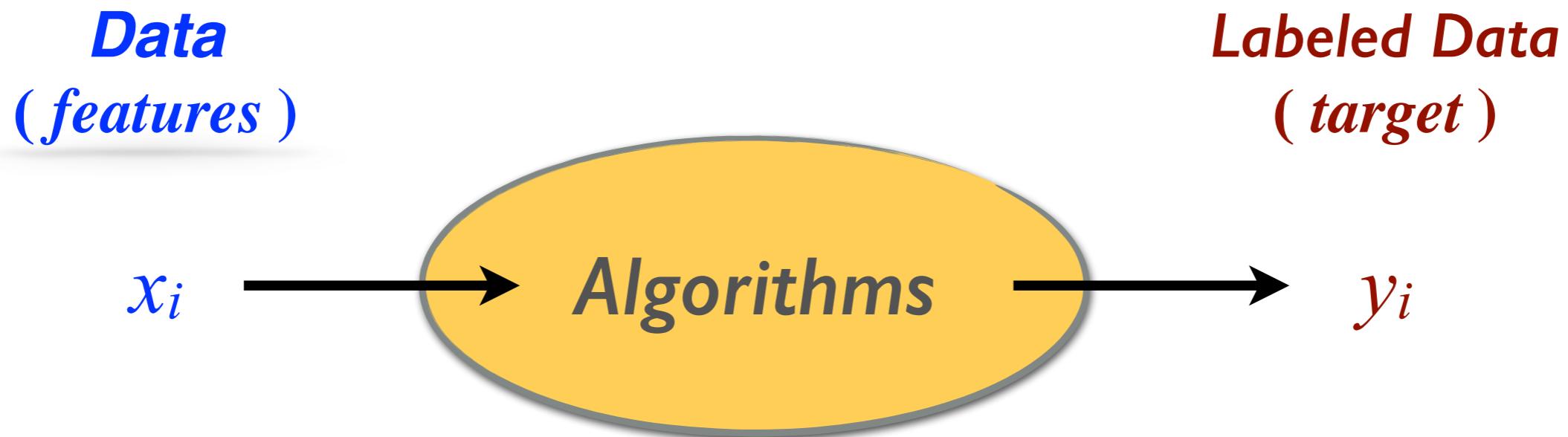
-> **Self-Adaptive Intelligent System** such as **AlphaGo** from **DeepMind**

Types of Machine-learning Algorithms

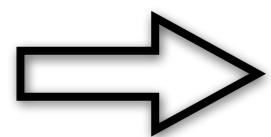
There are three different types of machine-learning algorithms for intelligent system development:

- Supervised machine-learning algorithms
- Unsupervised machine-learning algorithms
- Recommender systems

Supervised Learning — Predictive Model



$\{ x_i, y_i \}$: *training dataset* for algorithms

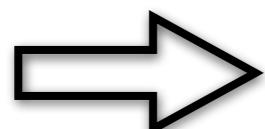


Supervised Learning

Unsupervised Learning — Descriptive Model



- $\{ x_i \mid x_1, x_2, x_3, \dots, x_n \}$ ↗ 1. *Pattern Discovery*
(Association Rules)
2. *Clustering*



Unsupervised Learning

Types of Machine-learning Algorithms (cont'd)

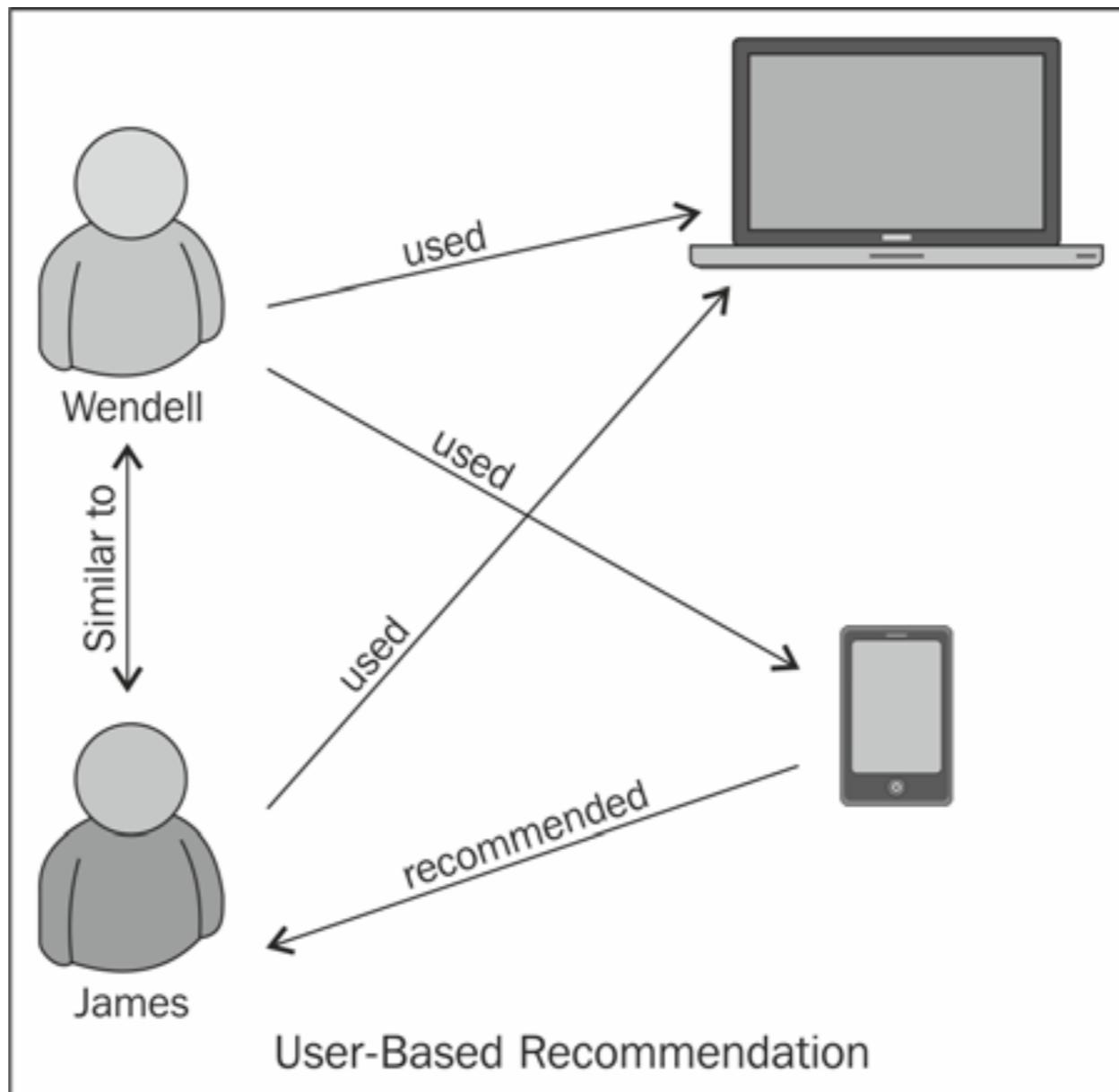
TYPE 3 : Recommendation algorithms

There are two different types of recommendations:

- User-based recommendations
- Item-based recommendations

TYPE 3 : Recommendation algorithms

User-based recommendations

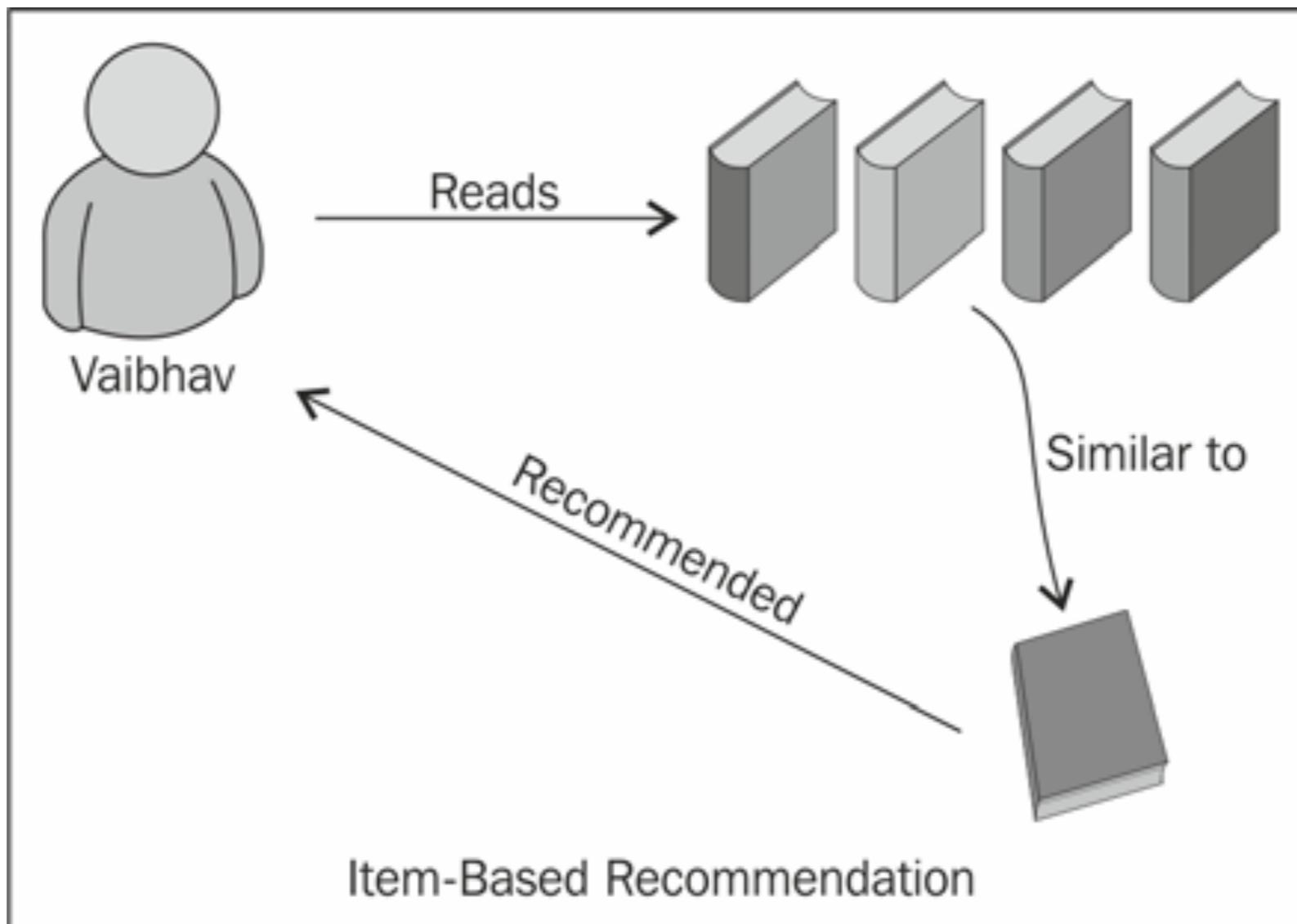


Types of Machine-learning Algorithms

(cont'd)

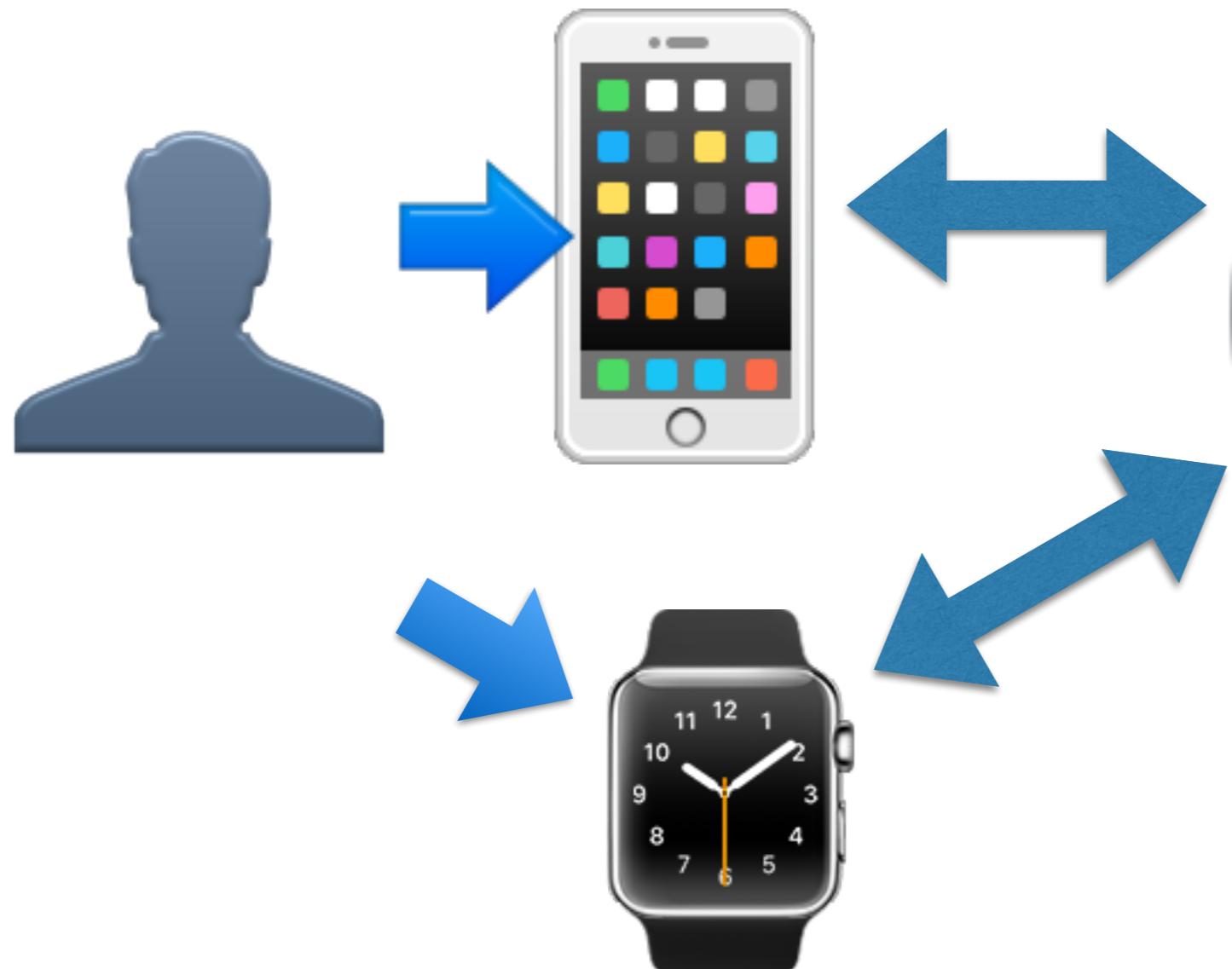
TYPE 3 : Recommendation algorithms

Item-based recommendations



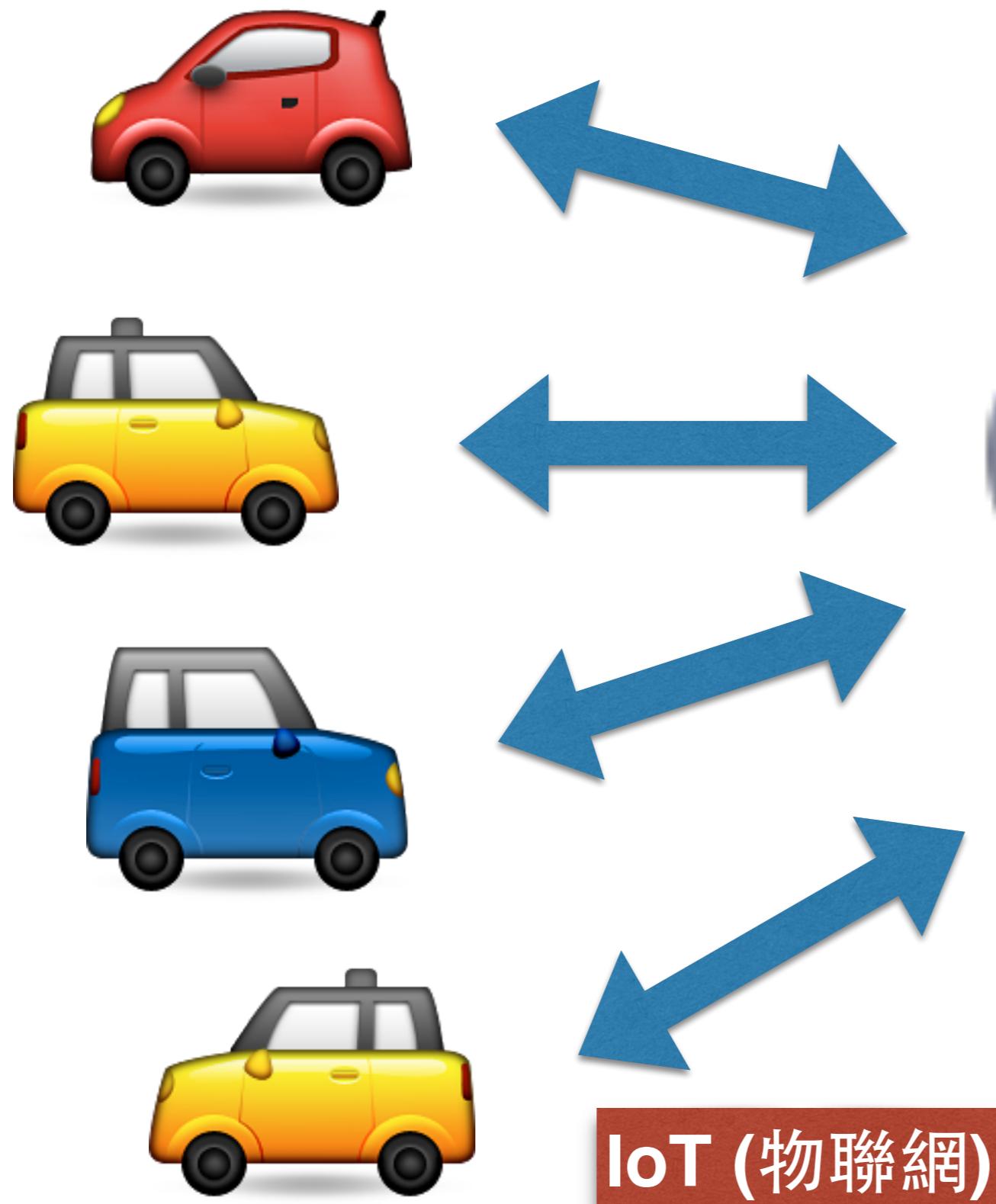
Big-Data Cloud System with Mobile Network

User: 中午吃什麼?
Big-Data Cloud System:
User: Yes / No



- Big Data 雲端決策支援系統**
- Expert Systems
 - 1. Rule-based
 - 2. Fuzzy Logic
 - 3. Bayesian Network
 -
 - AI : Machine Learning
 - (Data Mining)
 - Applications:
 - 生理數據分析 (心跳 脈搏)
 - 地理數據分析 (GPS)
 - 消費數據分析 (信用卡)
 - 查詢數據分析 (Google)
 - 商業決策支援系統 ...

Big Data Analytics in Engineering System : IoT Applications



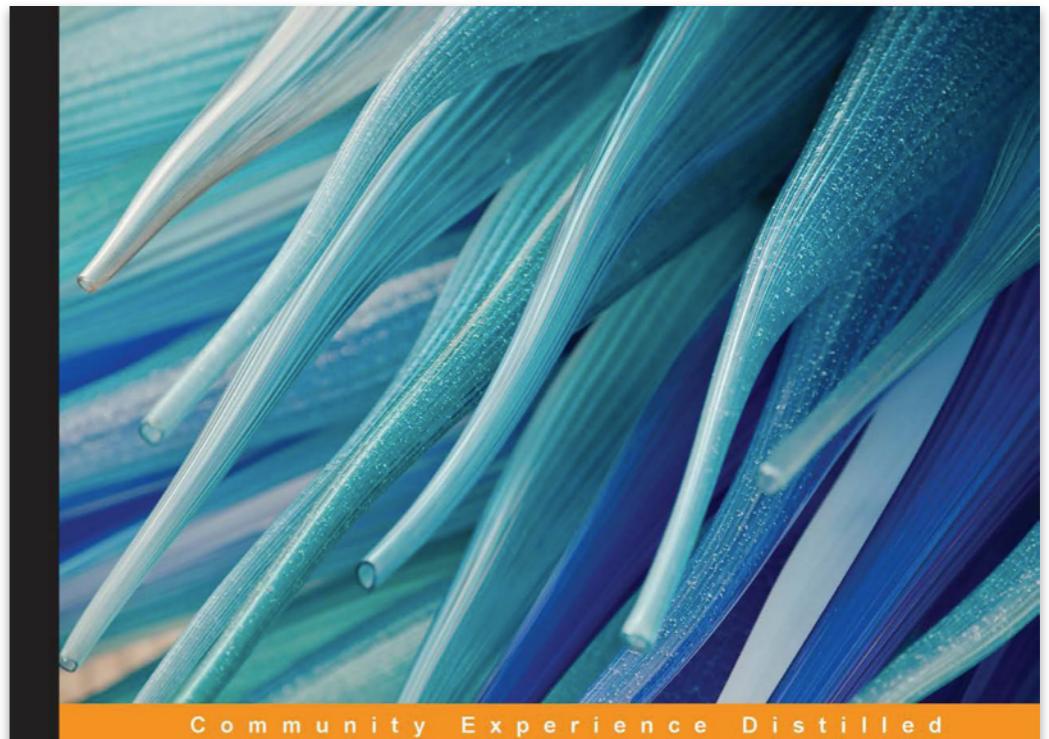
Big Data 雲端決策支援系統

- Expert Systems
 - 1. Rule-based
 - 2. Fuzzy Logic
 - 3. Bayesian Network
 -
- AI : Machine Learning
(Data Mining)
- IoT Applications:
 - 醫療照護 (護理站)
 - 工廠管理 (機台監控)
 - 環境工程
 - 無人駕駛汽車
 - 生物化學實驗分析
 - 運輸管理分析...

5. Machine Learning with R

Brett Lantz,
“Machine Learning with R,”
2nd ed., 2015.

<https://github.com/devharsh/Technical-eBooks/blob/master/Machine%20Learning%20with%20R,%202nd%20Edition.pdf>



Machine Learning with R *Second Edition*

Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R

Brett Lantz

[PACKT] open source*
PUBLISHING

6. Concluding Remarks

Data, Data & Data — 未來的獲利模式 (Google -> AI)

<< 商業週刊 1438期 (2015/6/8 ~ 6/14) >>

標題：“大數據的新生意經”

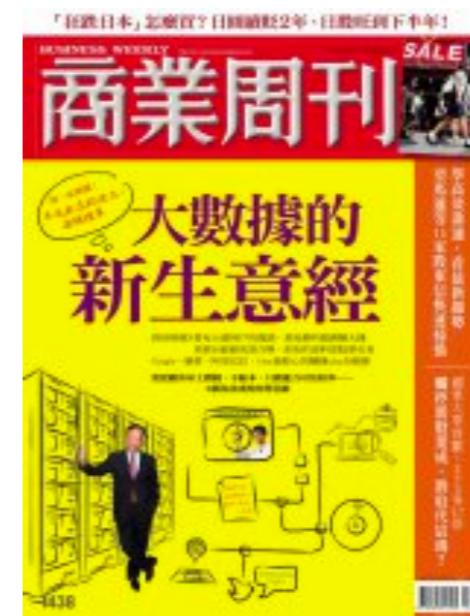
副標題 — “你一定要懂：

羊毛出在狗身上，由豬買單。”

(羊：消費者

狗：擁有大數據的企業

豬：花錢買大數據的企業)



- **AI + Big Data Analytics** — 智慧化的商業營運模式
- **Machine Learning (Data Mining)** — 資料科學家 的必備知識
- **R, Python and/or Scala** — 資料科學家 必備的程式設計能力

Thank you for listening !!

Q & A

