10. AI 代理人與機器學習理論基礎

< PART III Python 機器學習實務 >

OUTLINE

- 資料分析流程
- Raw Data (原始資料) 前置處理 — NumPy, Pandas
- 資料視覺化 — Matplotlib
- 特徵工程 (Feature Engineering)
- 機器學習 Training & Testing Pipeline — Scikit-Learn
- AI Agent & Learning Agent

原始資料
Raw Data

- Finance
- Med/Health
- Engineering
- … … …

1

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

傳統資料分析
vs.
Big Data 資料分析

2

傳統資料分析流程： R / Python / Scala

資料視覺化
(Data Visualization)

特徵向量擷取
(Feature Vector's Extraction)

資料分析方法：

- 機率模型
- 統計模型
- 資料探勘 (Data Mining)

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： PySpark, SparkR, Spark/Scala

Data Analytics Project Life Cycle (5 stages)

1

1. Identifying the problem

2. Designing data requirement

3. Preprocessing data

4. Performing analytics over data

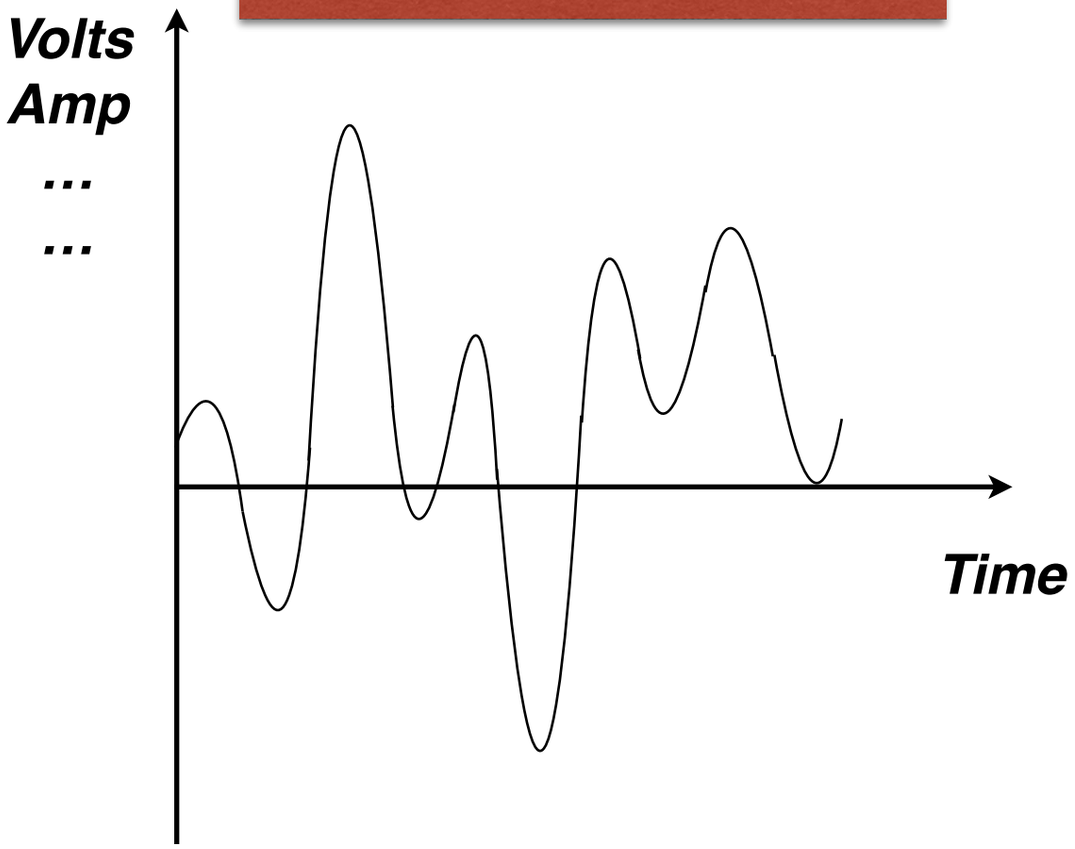
5. Visualizing data

Stage 1 ~ 2 :
Data Collection
~ 80% Work-load

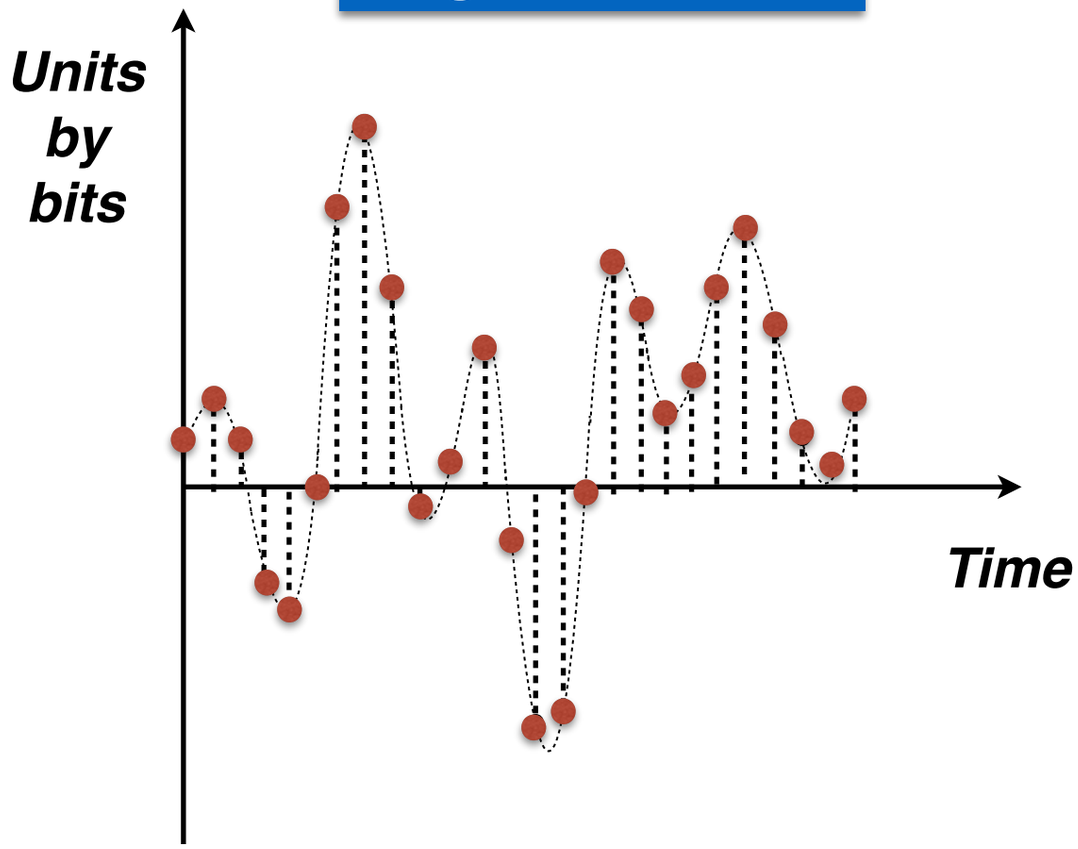
Stage 3 ~ 5 :
Data Analysis
~ 20% Work-load

1

Raw Data (原始資料)



Digitized Data



Data Acquisition
(數據擷取)

- Sampling Rate (取樣速率)
- Quantization (數據量化)

Analog Signal
(類比訊號)

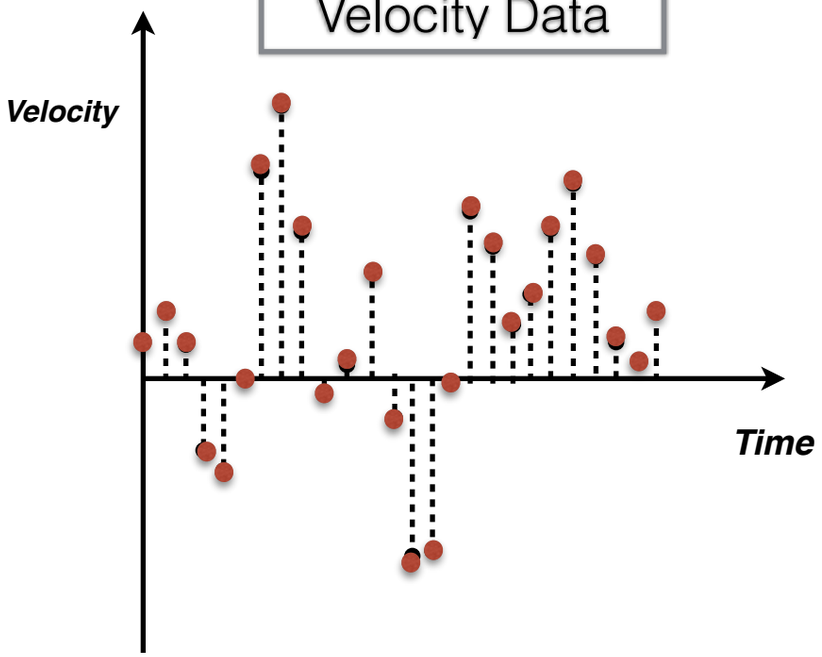
A/D C

Digital Signal
(數位訊號)

Digitized Dataset

無人自駕車、
無人機、
手機定位、
智慧型機器人、
智慧感測 ...

Velocity Data



GPS Data

Pitching Data

Rolling Data

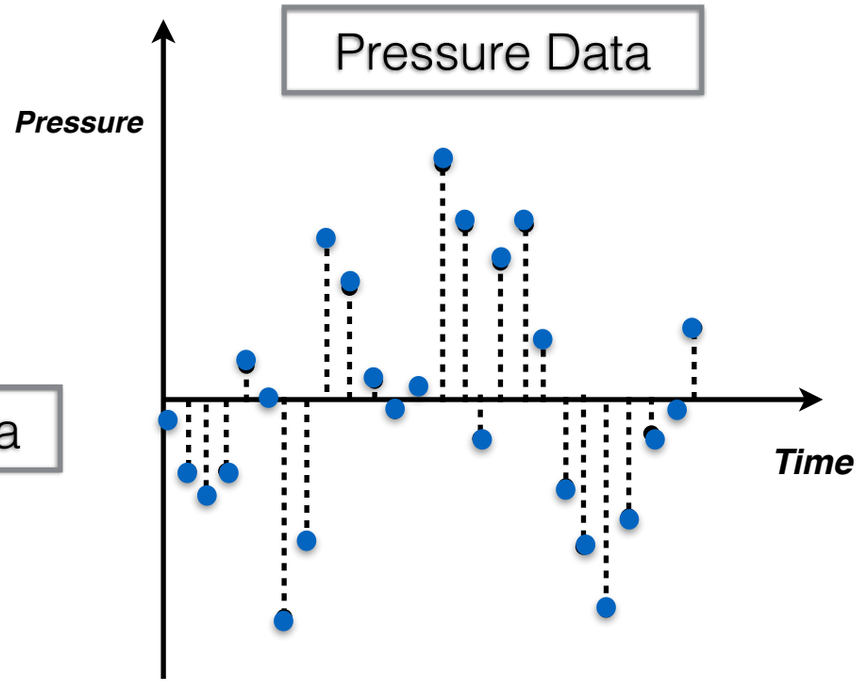
Yawing Data

Temperature Data

... ...

... ...

Pressure Data



1 Data Pre-processing (資料前置處理)

需要領域知識

Data Tables

Time	GPS	Velocity	Pitching	... ...
t_1	...	...	...	...
t_2	...	...	...	...
t_3	...	...	...	...
...	...	...	...	...
...	...	...	...	...

1 Data Pre-processing：產線需建立 Data Tables

[需要用到的 Python 技術]
NumPy & Pandas

[One Example]： biopsy data

ID	area	shape	texture	…
id1	…	…	…	…
id2	…	…	…	…
id3	…	…	…	…
…	…	…	…	…
…	…	…	…	…

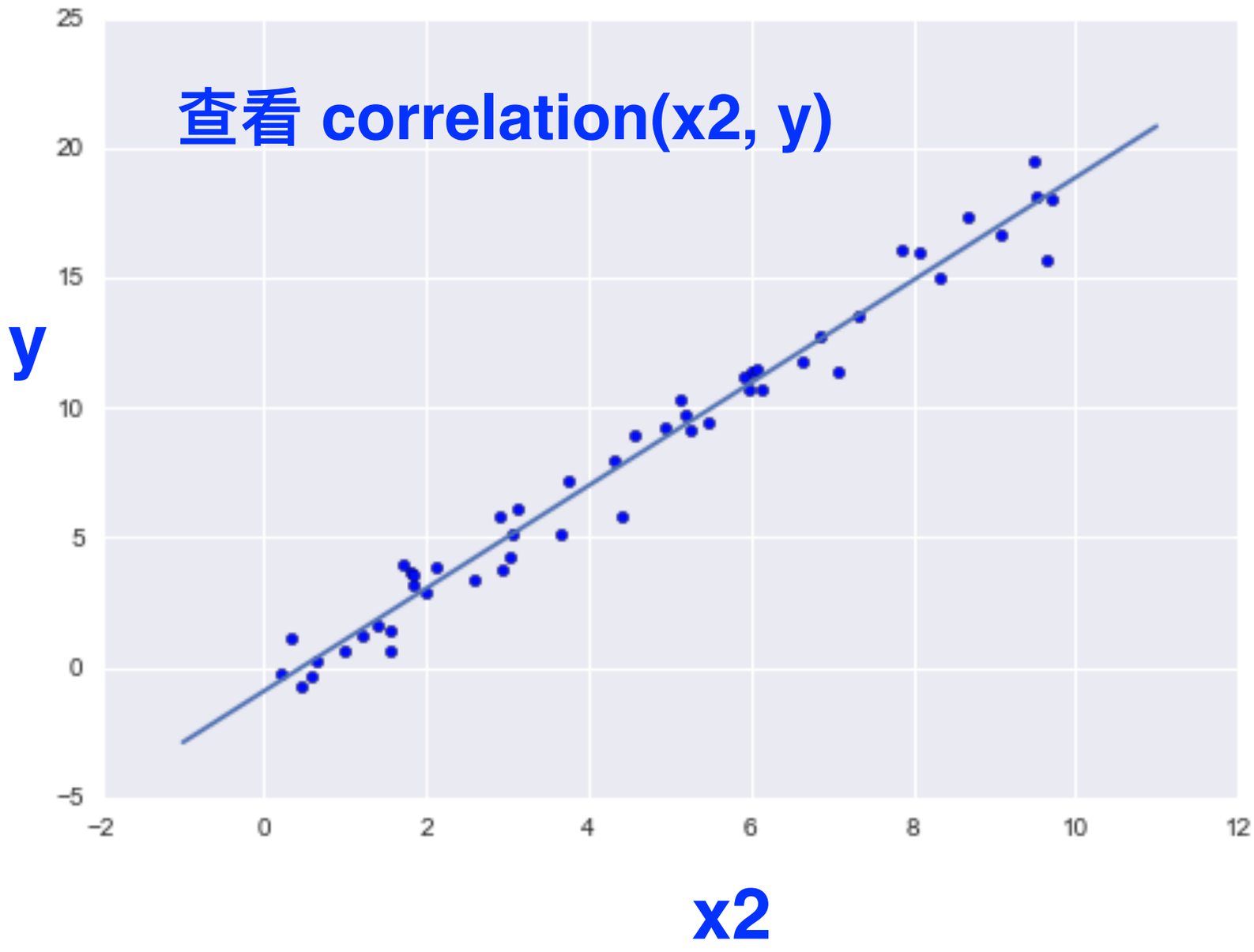
[Another Example]： AAPL 股票

Date	Open	High	Low	Closed	…
d_1	…	…	…	…	…
d_2	…	…	…	…	…
d_3	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…

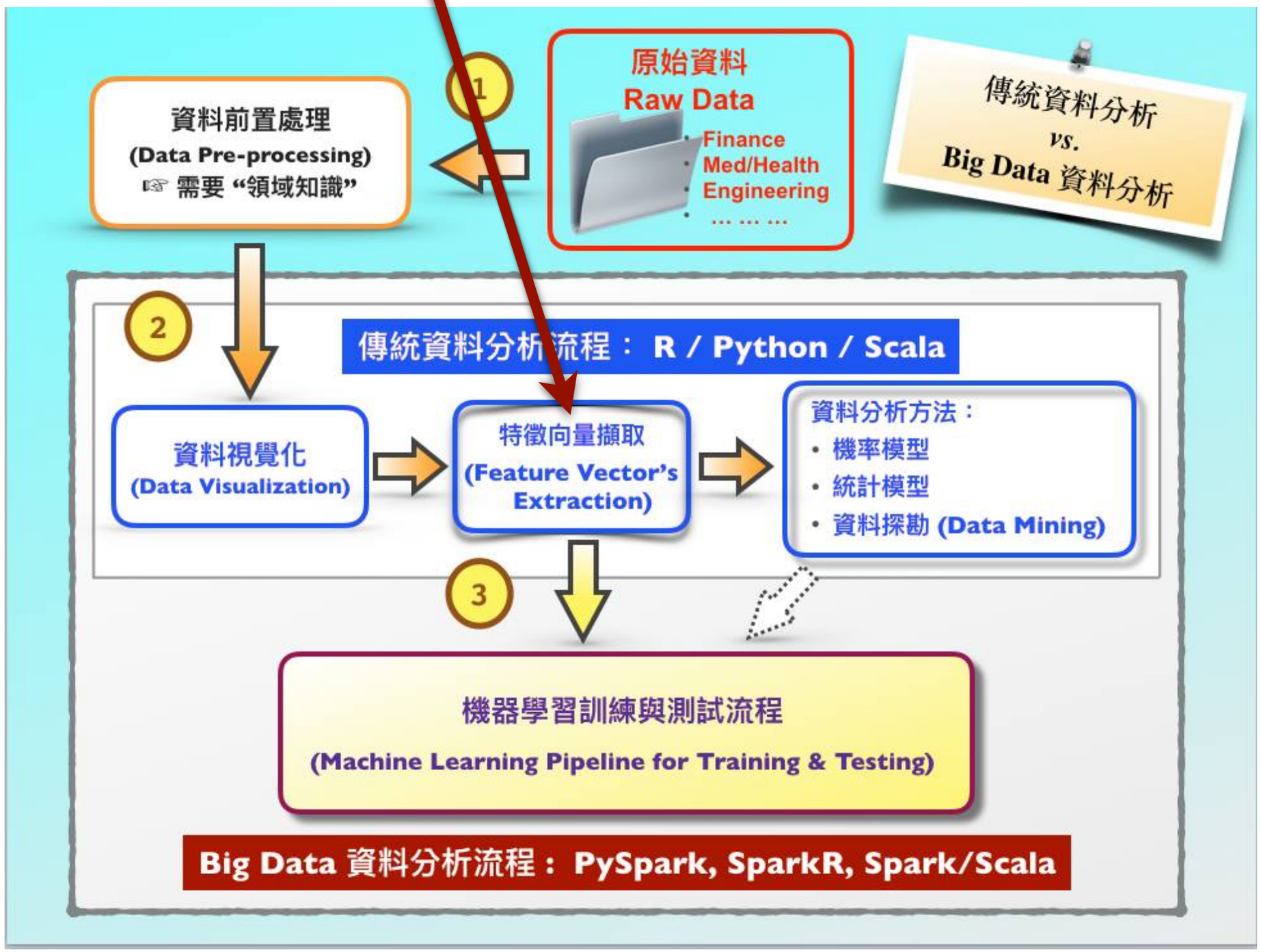
2 Data Visualization：從 Data Tables 繪圖

[需要用到的 Python 技術]： Matplotlib

No.	x1	x2	x3	y
1	…	…	…	…
2	…	…	…	…
3	…	…	…	…
4	…	…	…	…
…	…	…	…	…



2 Feature Engineering：決定 Feature Variables



2 Feature Engineering：決定 Feature Variables

Supervized Learning — Predictive Model

Data
(features)

Labeled Data
(target)

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

⇨ **Supervized Learning**

=> Machine Learning Training & Testing Pipeline

3 Machine Learning Training & Testing Pipeline

Datasets : Feature + Labeled, (x_i , y)

Training Dataset, *e.g.* 75% ~ 80%

Transformation
(Normalization or Standardization ?)

Algorithm(s)

Evaluation

Testing Dataset, *e.g.* 20% ~ 25%

Training Model

Prediction

3 Machine Learning Training & Testing Pipeline

[需要用到的 Python 技術] ： Scikit-Learn

Q： 為什麼 Feature Dataset, x_i 要拆解成 75% (or 80%) : 25% (or 20%) ？
可不可以拆解成其他比例呢？ 例如：60% : 40%

Q： 如何決定該選擇 Normalization 或 Standardization ？

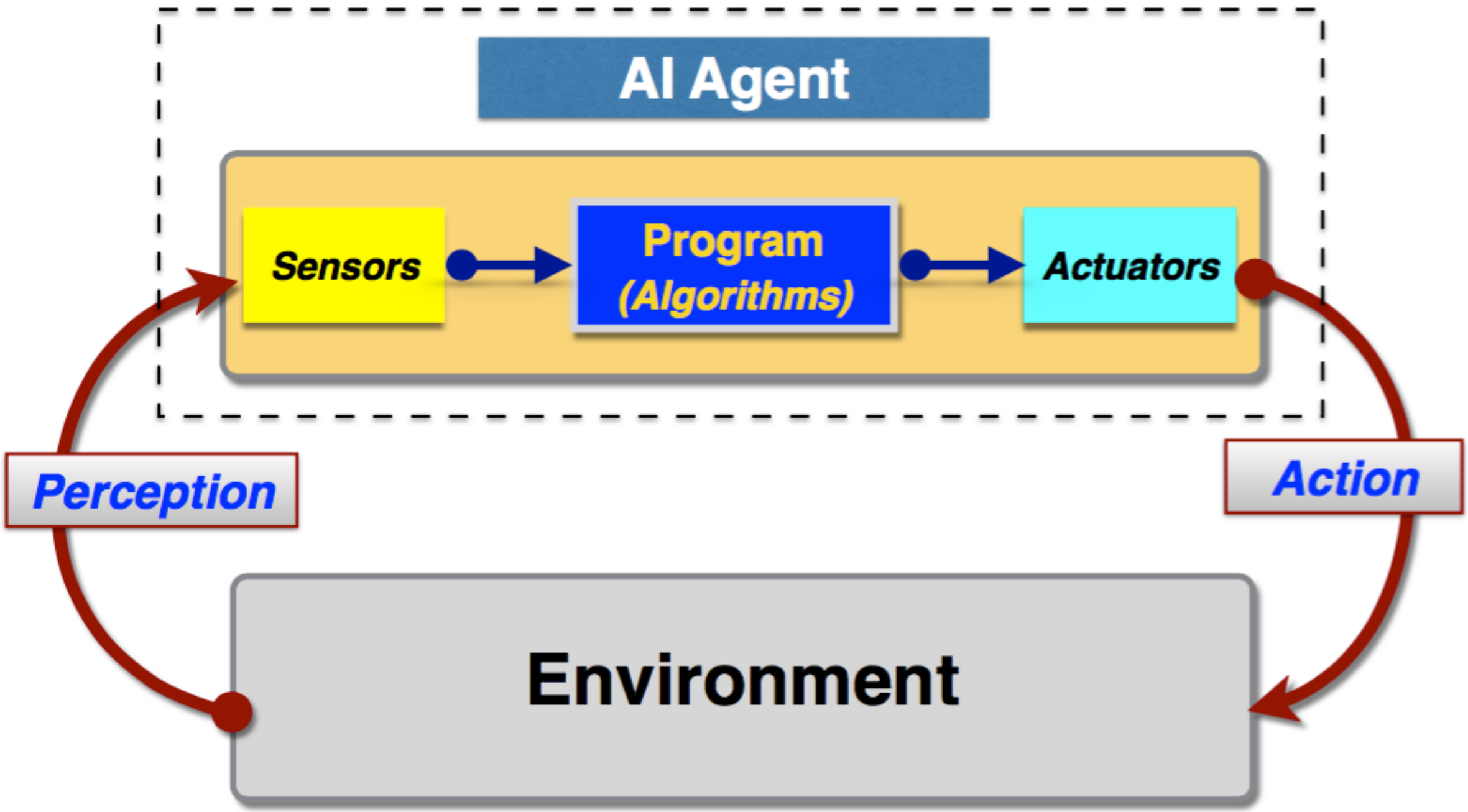
Q： 如何在 Training 階段 選擇 Algorithms ？

Q： 如何在 Training 階段，進行 Evaluation (評估)？ [labeled data]

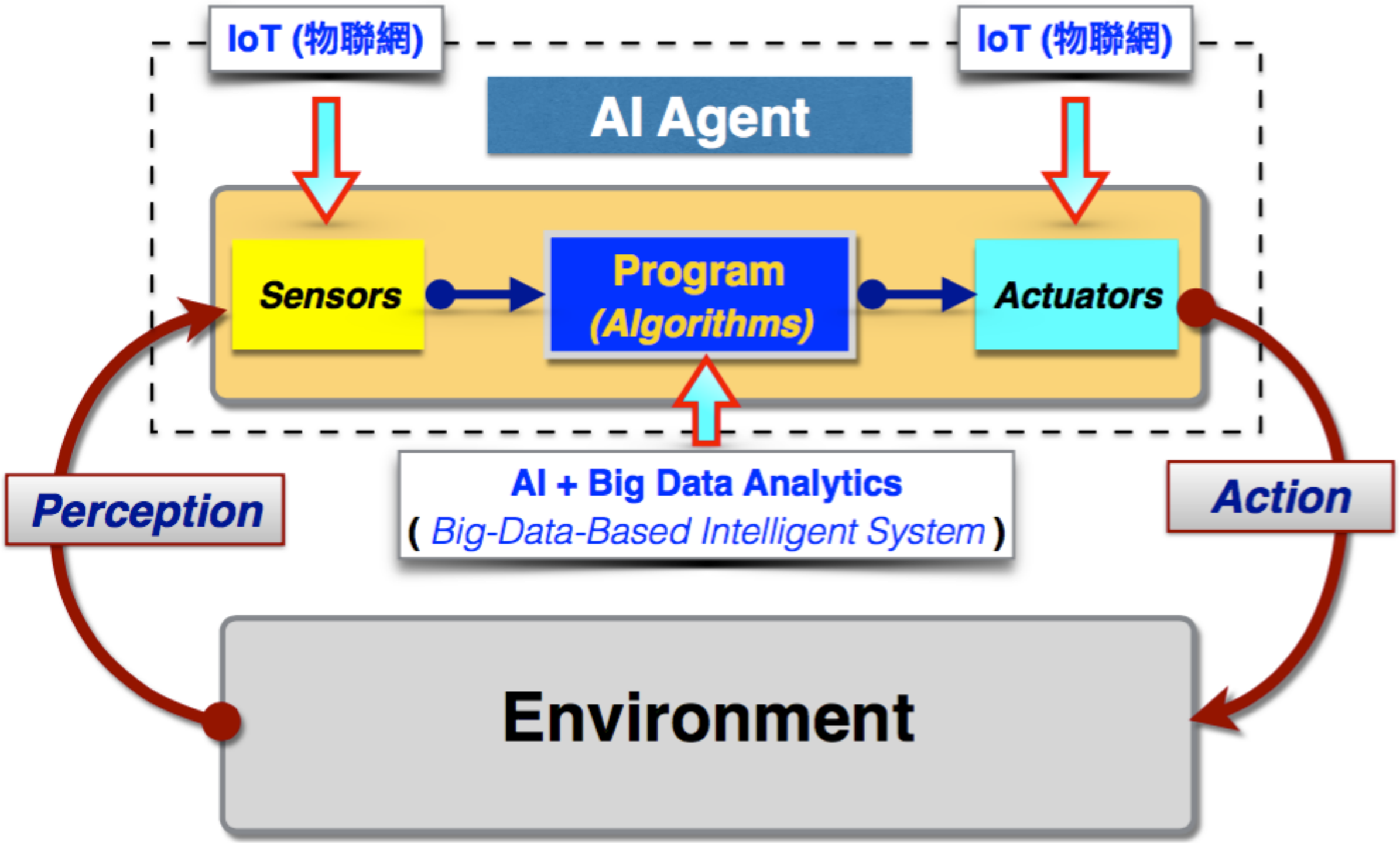
Q： 如果 Training Model 在 Testing 階段，預測結果表現 “很好” 或 “不好”，該如何處理呢？

Risk Analysis： Test Result => Baseline => Strategy
Risk Management => Execution

AI Agent


AI Agent
Sensors
Program
(Algorithms)
Actuators
Perception
Action
Environment

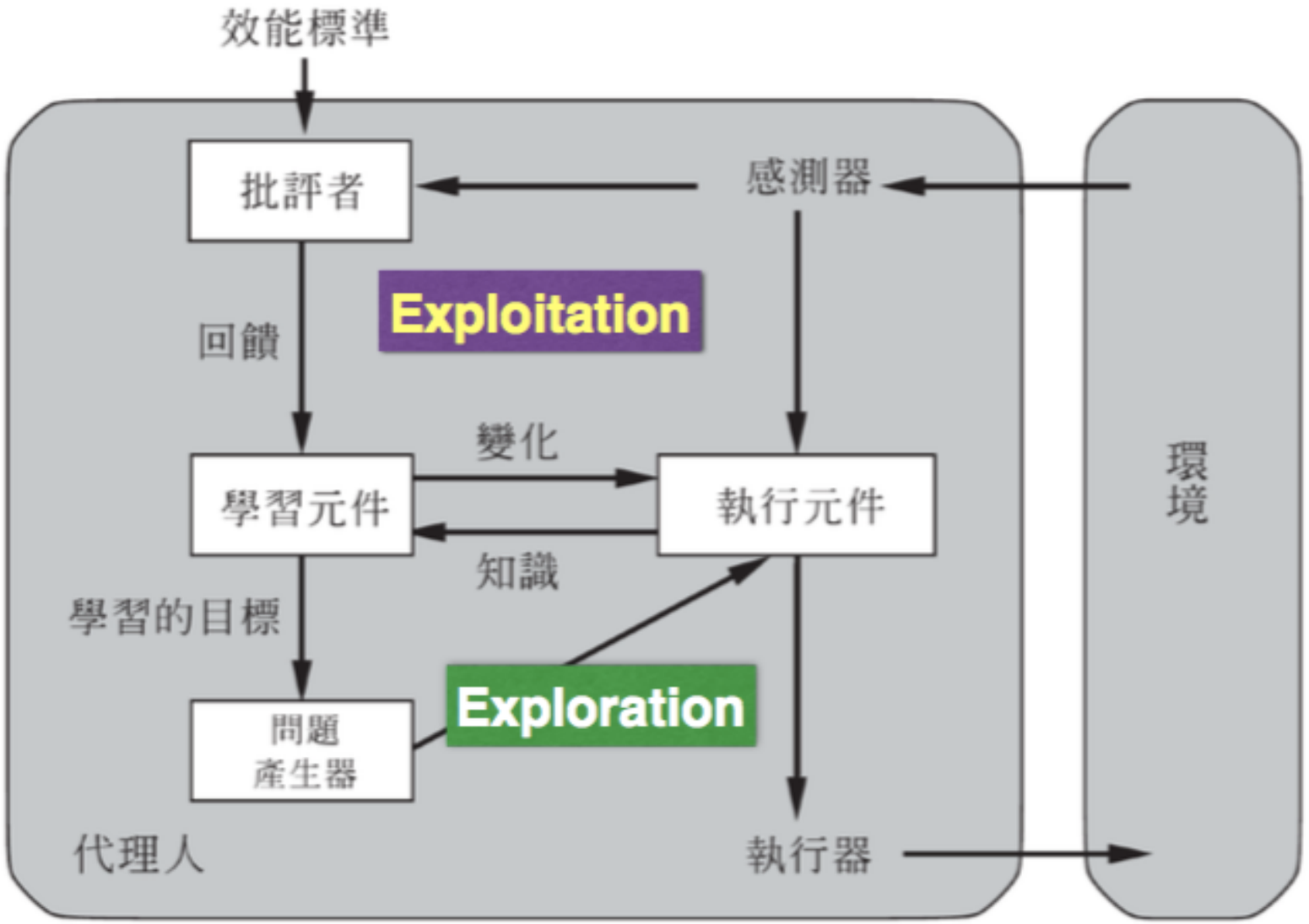
AI Agent for Industry 4.0 Solutions



一般的學習型代理人 *Exploitation vs. Exploration*

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



一般的學習型代理人

REFERENCE

Jake VanderPlas, "***Python Data Science Handbook***", **2016, O'Reily.**

https://jakevdp.github.io/PythonDataScienceHandbook/

Code from Github :

https://github.com/jakevdp/PythonDataScienceHandbook