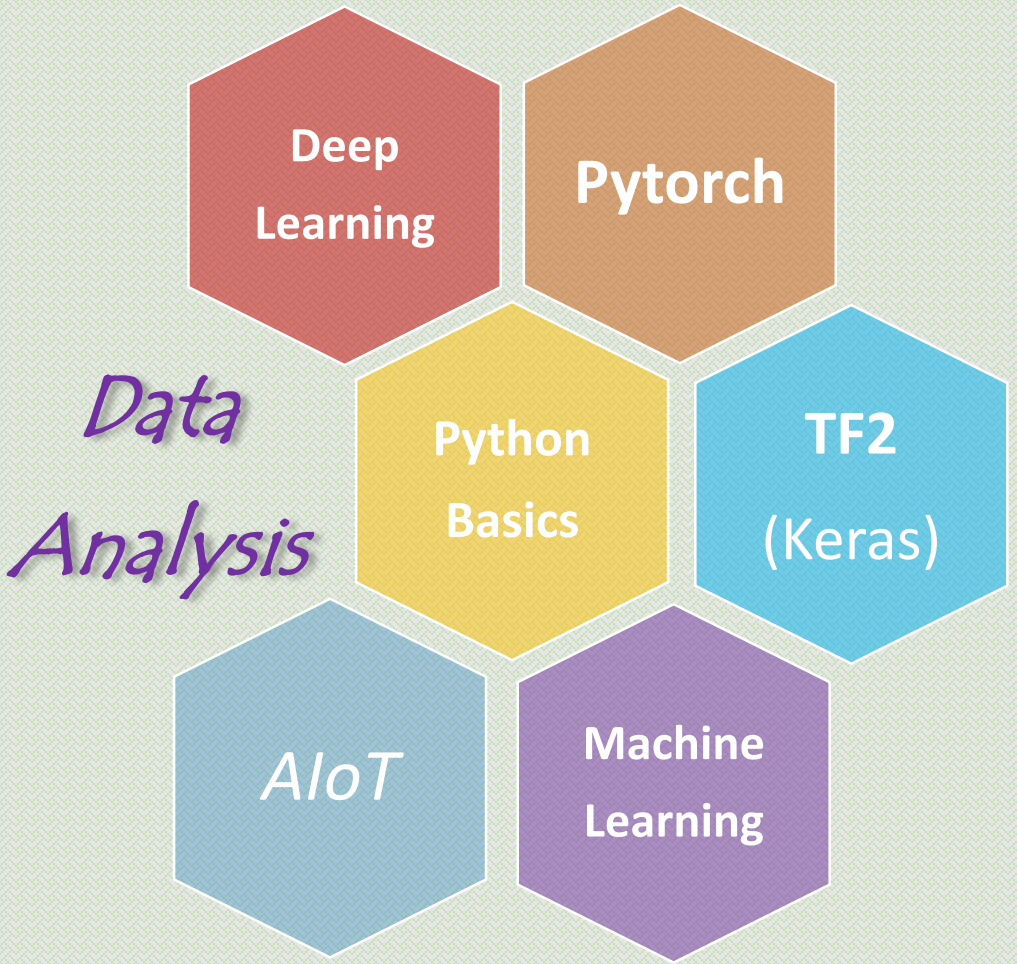
< PART IV>

Feature Engineering

(特徵工程)



胡中興 (*C. Alex Hu, PhD*)

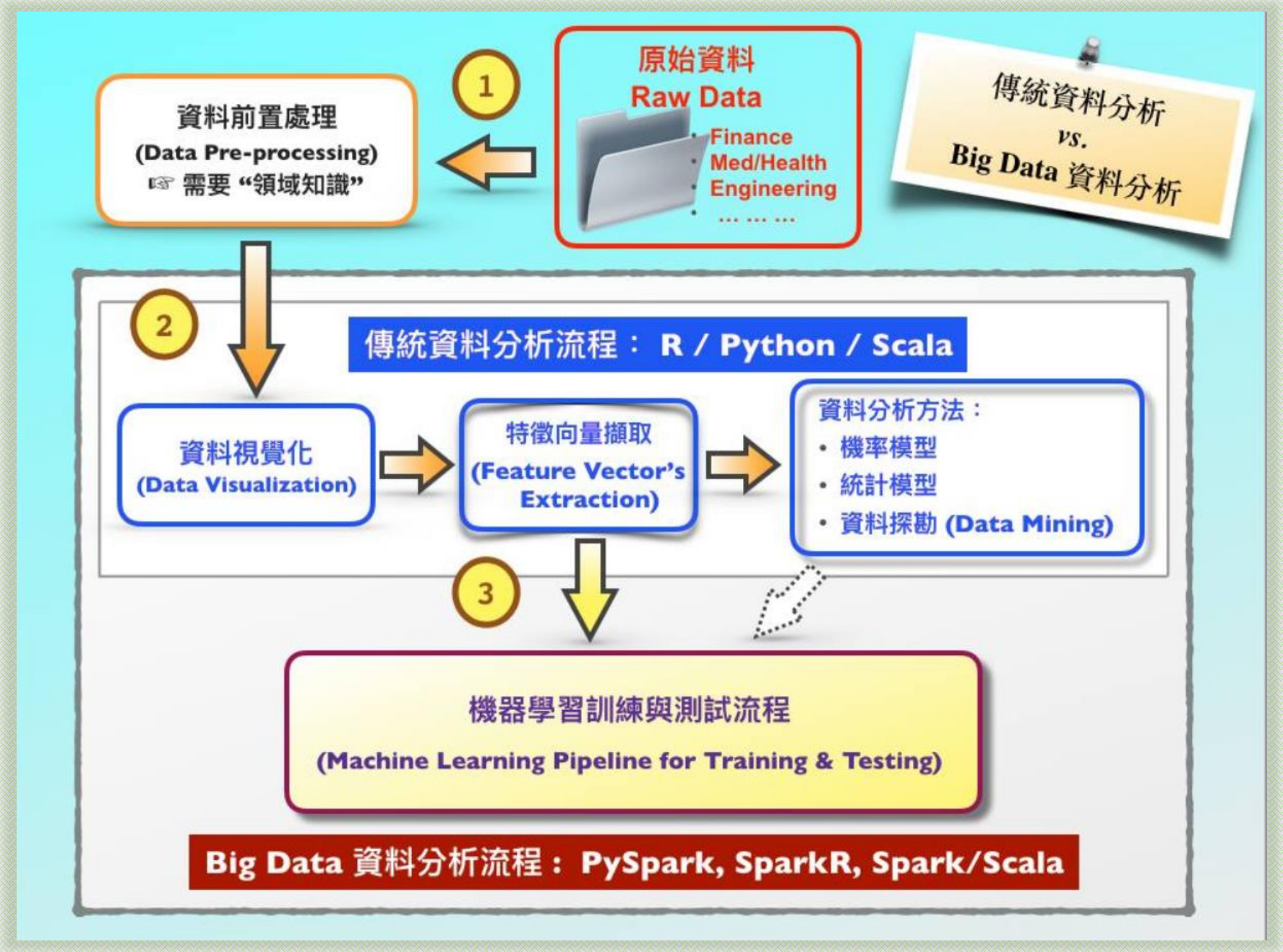
Courses on "Python Data Analysis (Python 資料分析)"

OUTLINE

- Data Analysis & Machine Learning Pipeline
- Feature Engineering
  - ❖ Feature Extraction
  - ❖ Feature Selection
- Scikit-Learn Workshop for Feature Engineering

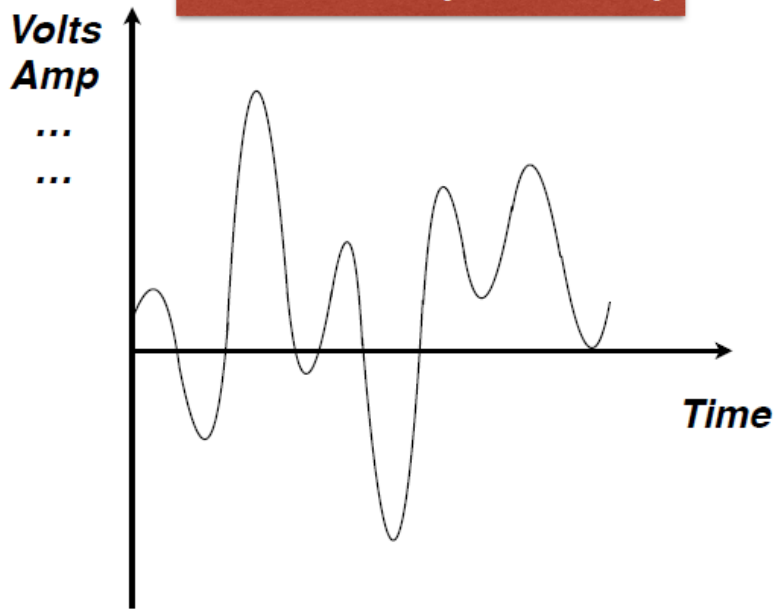
Data Analysis & Machine Learning Pipeline

（資料分析與機器學習流程）

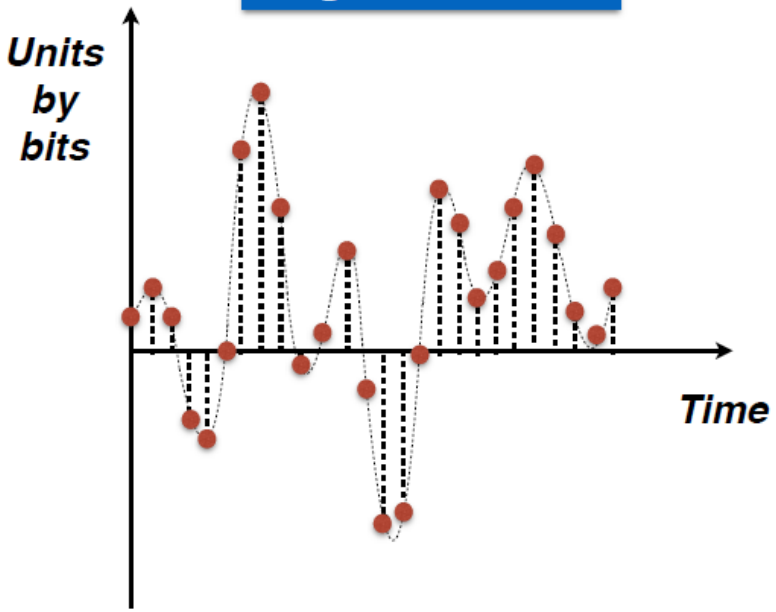


1

Raw Data (原始資料)



Digitized Data



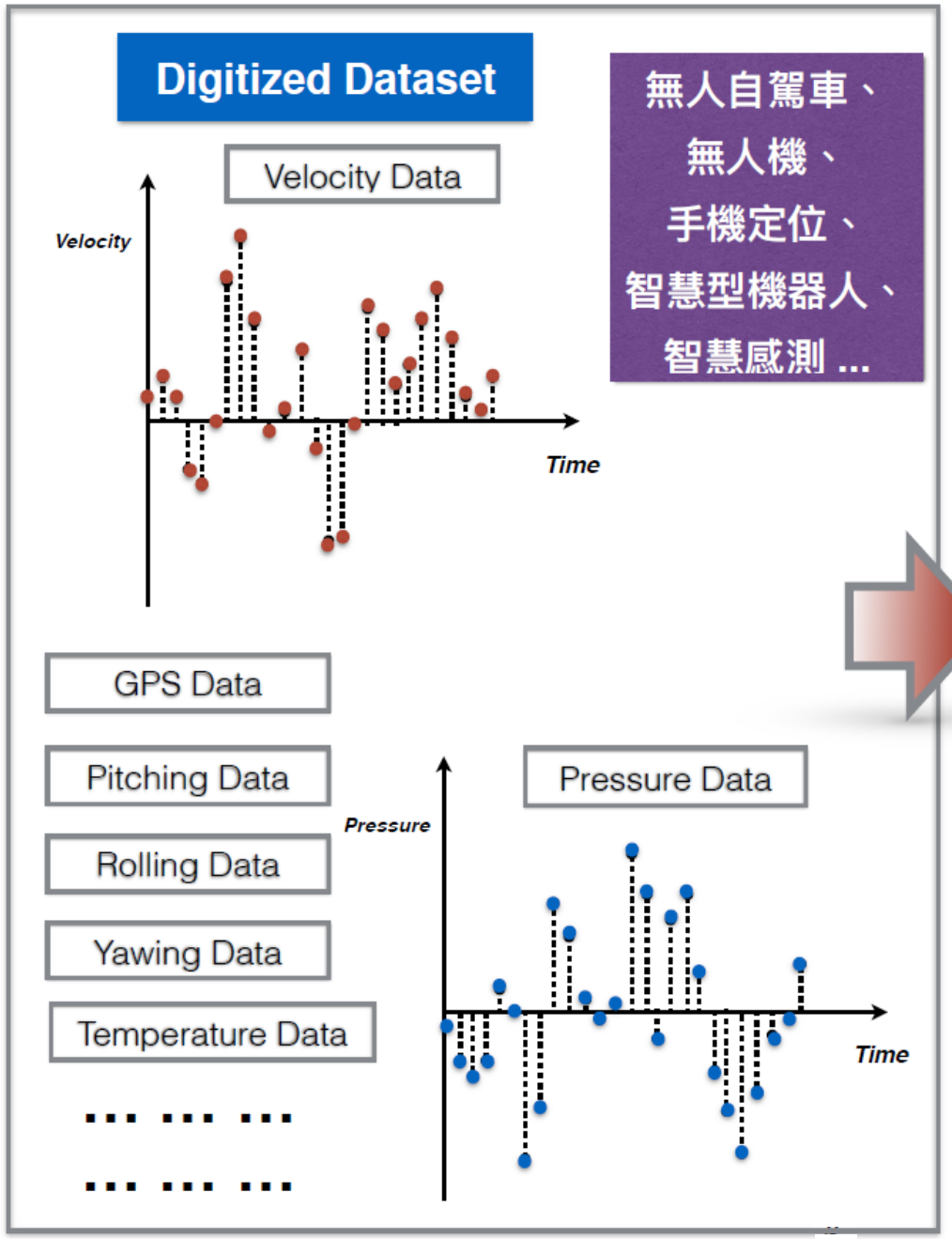
Data Acquisition
(數據擷取)

- **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)



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	…	…	…	..
…	…	…	…	..
…	…	…	…	..

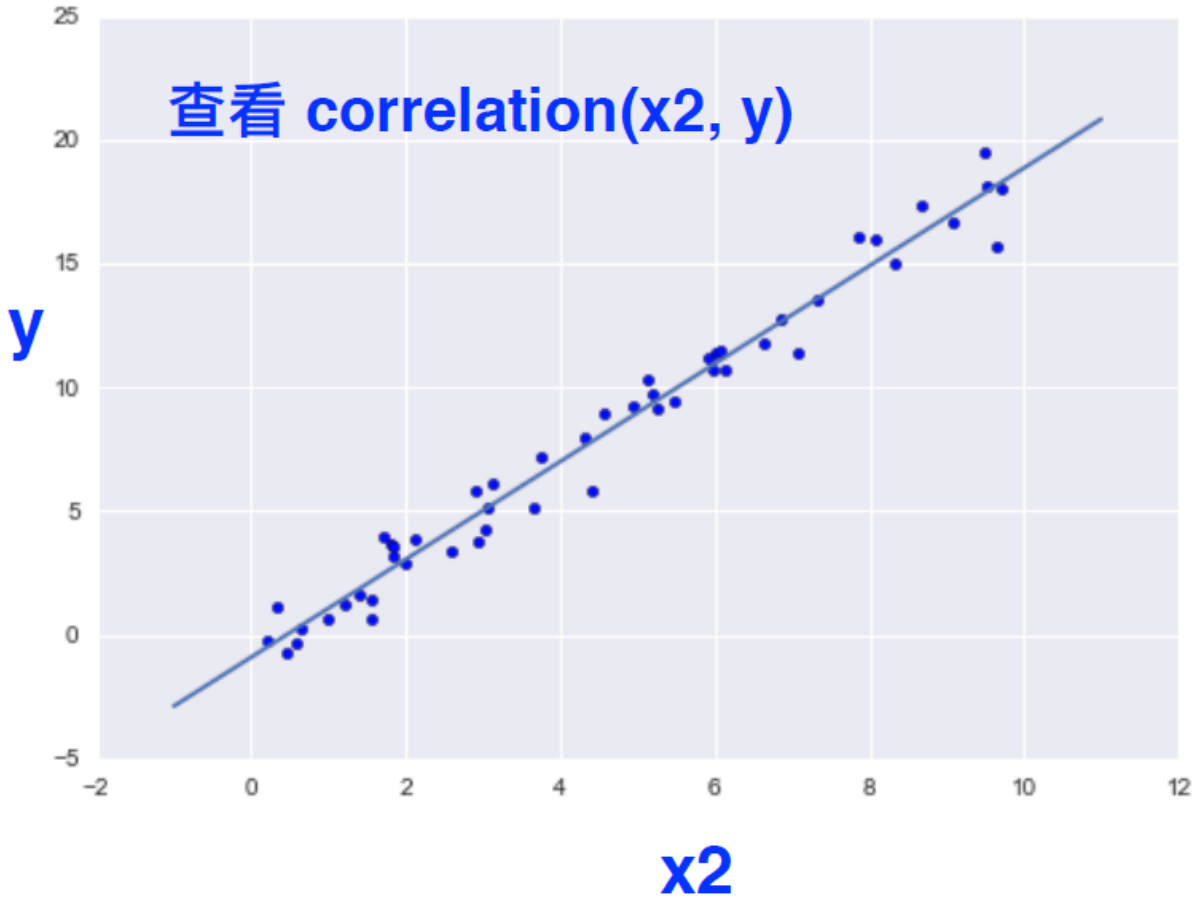
[Another Example] : AAPL 股票

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

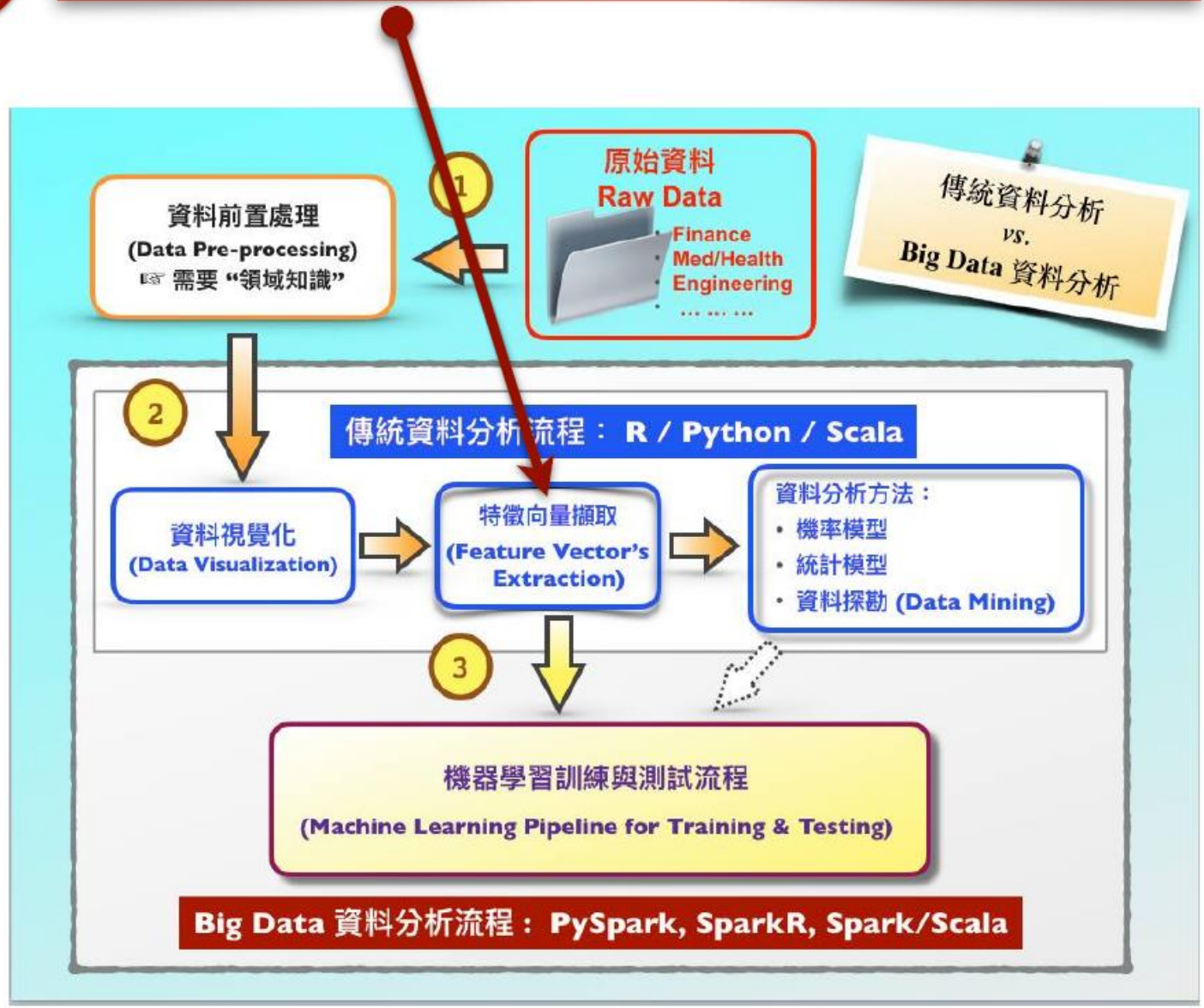
2 Data Visualization : 從 Data Tables 繪圖

[需要用到的 Python 技術] ： Matplotlib

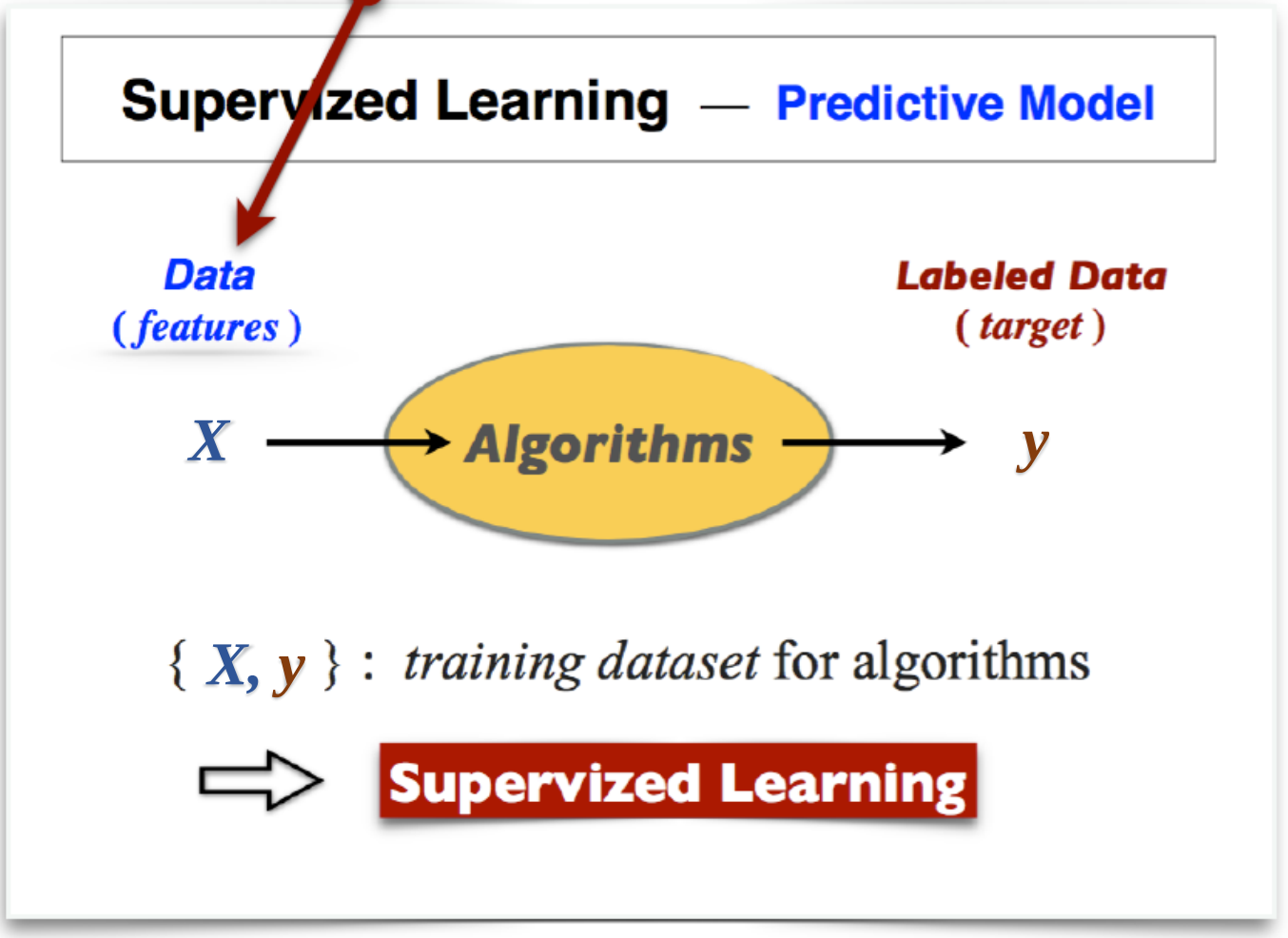
No.	x1	x2	x3	y
1	…	…	…	…
2	…	…	…	…
3	…	…	…	…
4	…	…	…	…
…	…	…	…	…



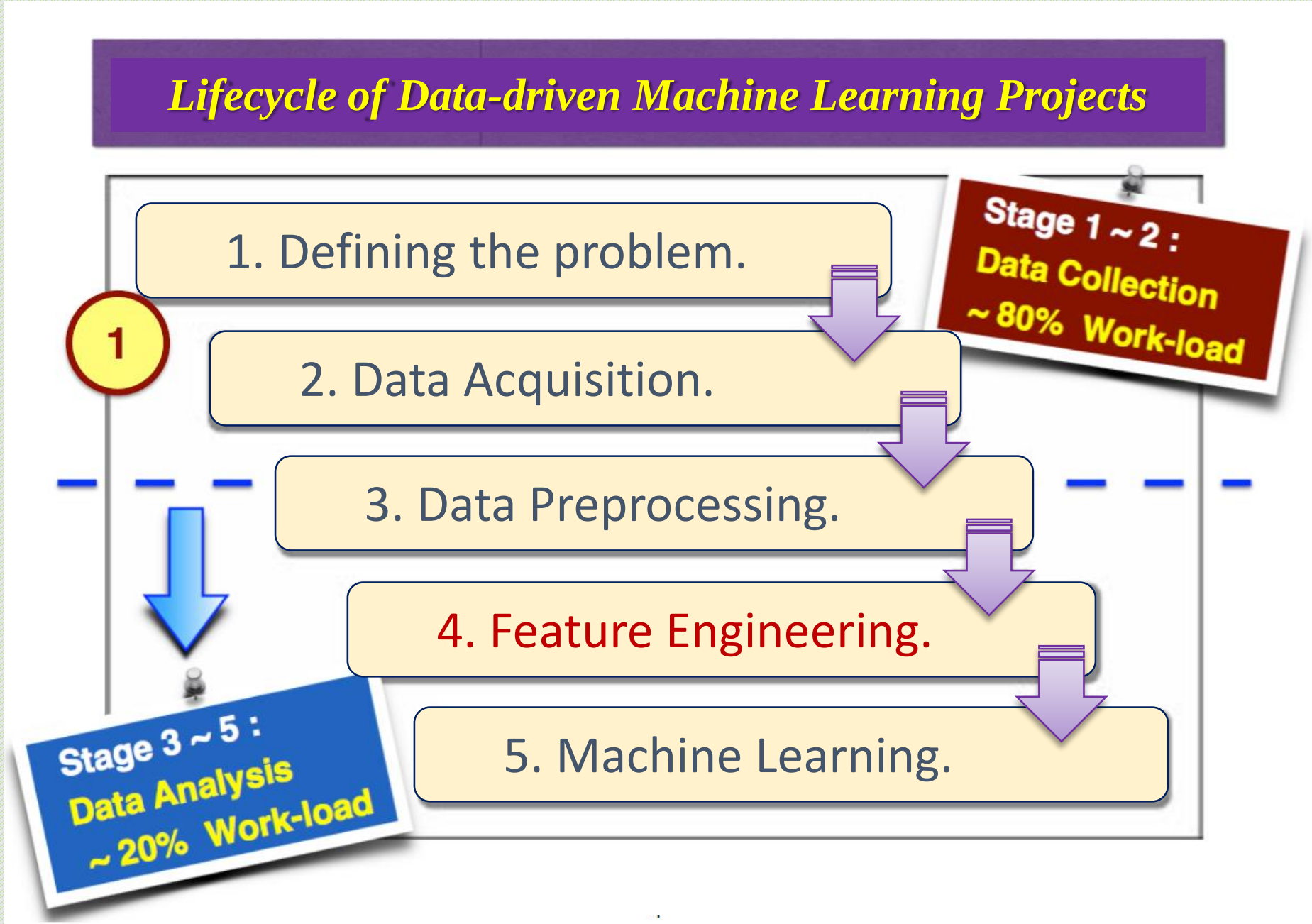
2 Feature Engineering：決定 Feature Variables



2 Feature Engineering : 決定 Feature Variables



=> Machine Learning Training & Testing Pipeline


Lifecycle of Data-driven Machine Learning Projects
1. Defining the problem.
2. Data Acquisition.
3. Data Preprocessing.
4. Feature Engineering.
5. Machine Learning.
1
Stage 1 ~ 2 :
Data Collection
~ 80% Work-load
Stage 3 ~ 5 :
Data Analysis
~ 20% Work-load

Feature Engineering

Data Preprocessing

- Vectorization
- Normalization
- Missing Values
- … … …
- Data Visualization

Feature Engineering

- Feature Extraction
- Feature Selection
  - Feature Importance
  - Dimensionality Reduction
  - … …

Machine Learning

- Supervised Learning
- Unsupervised Learning

STEP 1

Feature Extraction

– Obtain a pool of appropriate features from the pre-processed data for *feature selection*.

STEP 2

Feature Selection

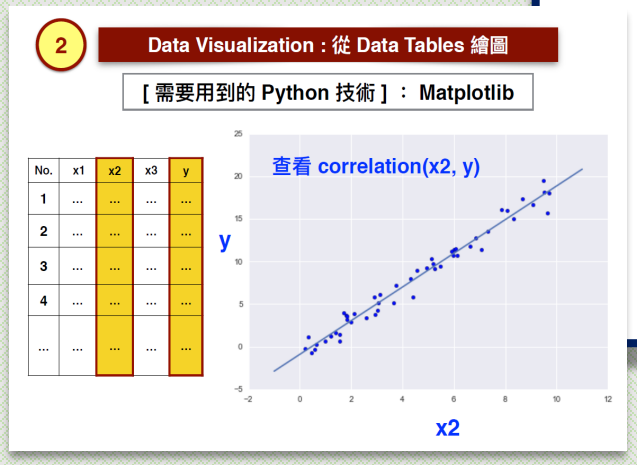
– Select a set of useful features out of the feature pool for *Machine Learning*.

Feature Importance

Dimensionality Reduction

Data Preprocessing

- Vectorization
- Normalization
- Missing Values
- … … …
- *Data Visualization*



Feature Engineering

- Feature Extraction
- Feature Selection

STEP 1

Feature Extraction

– Obtain a pool of appropriate features from the pre-processed data for *feature selection*.

STEP 1 : ***Feature Extraction***

by ***Data Visualization***

- *Pair plots*
- *Correlation matrix* (*by* heatmap)

[Example] : Iris dataset

Feature Engineering

- Feature Extraction
- Feature Selection

Machine Learning

Feature Selection

– Select a set of useful features out of the feature pool for *Machine Learning*.

STEP 2 *: Feature Selection –* ***Scikit-learn***

1. *Removing features with low variance*
2. *Univariate feature selection*
3. *Recursive feature elimination*
4. *Feature selection using* SelectFromModel
   - *L1-based feature selection*
   - *Feature Importance*

   [Example] : Mobile Price dataset
5. *Dimensionality Reduction*

   [Example] : Digit Recognition dataset

Scikit-learn Workshop

Scikit-Learn Workshop

for

Feature Engineering

CONTENT:

Iris dataset

Mobile Price

Digit Recognition