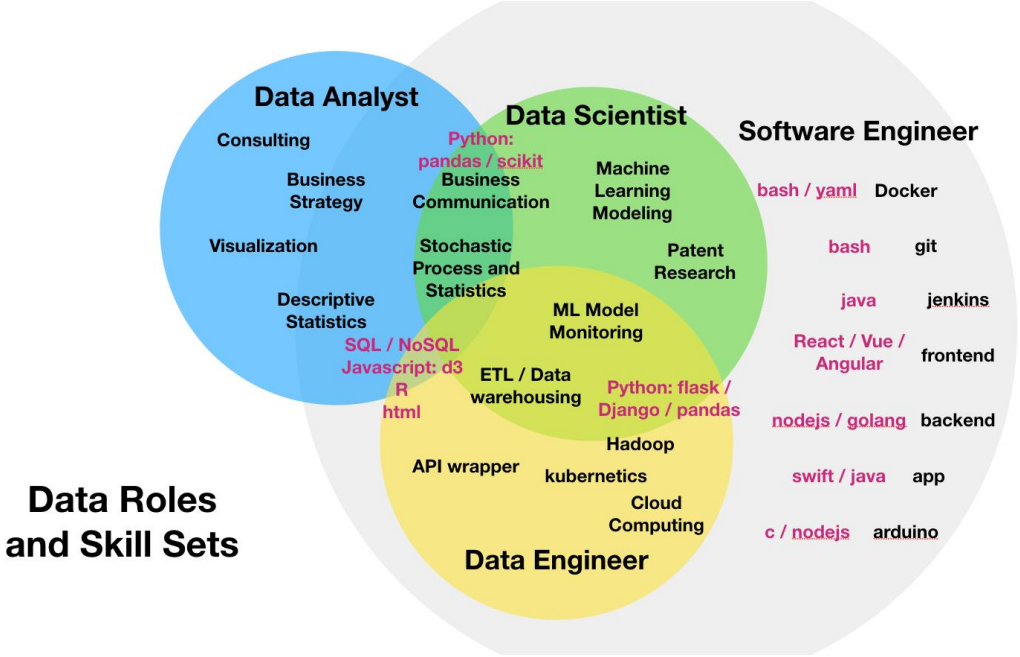
# 如何進行一個AI project?

黃志勝
義隆電子 人工智慧研發部
國立陽明交通大學 AI學院 合聘助理教授

# AI工程師和資料科學家


圖片來源:李弘毅教授投影片

# AI工程師和資料科學家




https://vocus.cc/dnsc/5c0237c9fd89780001e51d88


繪圖：陳威安

https://aiacademy.tw/start-like-this/

# AI project?

**<u>AI projects</u>**: 資料科學、機器學習、深度學習、強化學習

- 資料科學: 離職預測、庫存分析、商品良率分析…等。
- 機器學習、深度學習: 車流偵測、人臉辨識、假/指紋辨識…等。
- 強化學習: 空調控制、紅綠燈控制…等。

# AI 學習方法

- What is machine learning?

Machine learning algorithms build a "**mathematical model**" based on sample data, known as "**training data**", in order to make predictions or decisions without being explicitly programmed to perform the task. [https://en.wikipedia.org/wiki/Machine_learning]

What is the "KEY" in machine learning?

ANS: Mathematics and Data.

# How to start a machine/deep learning application

- 1. Tasks/Applications definition.
- 2. Data (collection, labeling)
- 3. Learning Model
- 4. Evaluation model

# How to start a machine/deep learning application

- 1. Tasks/Applications definition. Marketing, PM
- 2. Data (collection, labeling) 資料工程師
- 3. Learning Model 機器學習AI工程師
- 4. Evaluation model 機器學習AI工程師

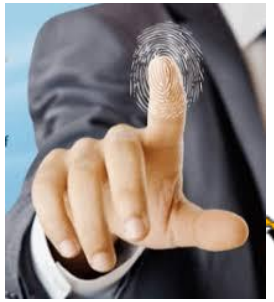

https://aiacademy.tw/start-like-this/

# Tasks/Applications definition

Please do **one right thing** and **one thing right**.

1. Specific Problem (right thing):
   What problem do you want to solve?

- 離職預測、庫存分析、商品良率分析…等。
- 車流偵測、人臉辨識、假/指紋辨識…等。
- 空調控制、紅綠燈控制…等。



# Tasks/Applications definition

Please do **one right thing** and **one thing right**.

2. Right Data (thing right):

   What is useful data? More information more better?

指紋辨識

# Data

- 為什麼Data很重要?
- 假設我們要做的分類問題是「男生和女生辨識」
- 特徵資料: 我們收集了三個資訊(居住地、身高和體重)

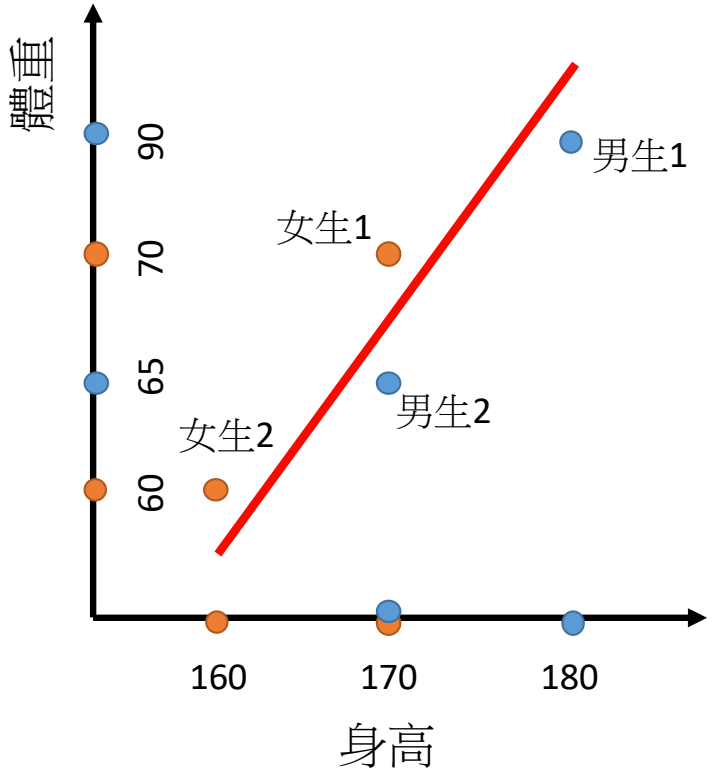
| | 居住地 | 身高 | 體重 |
|---|---|---|---|
| 男生1 | 新北市 | 180 | 90 |
| 女生1 | 新北市 | 170 | 70 |
| 男生2 | 新北市 | 170 | 65 |
| 女生2 | 新竹市 | 160 | 60 |

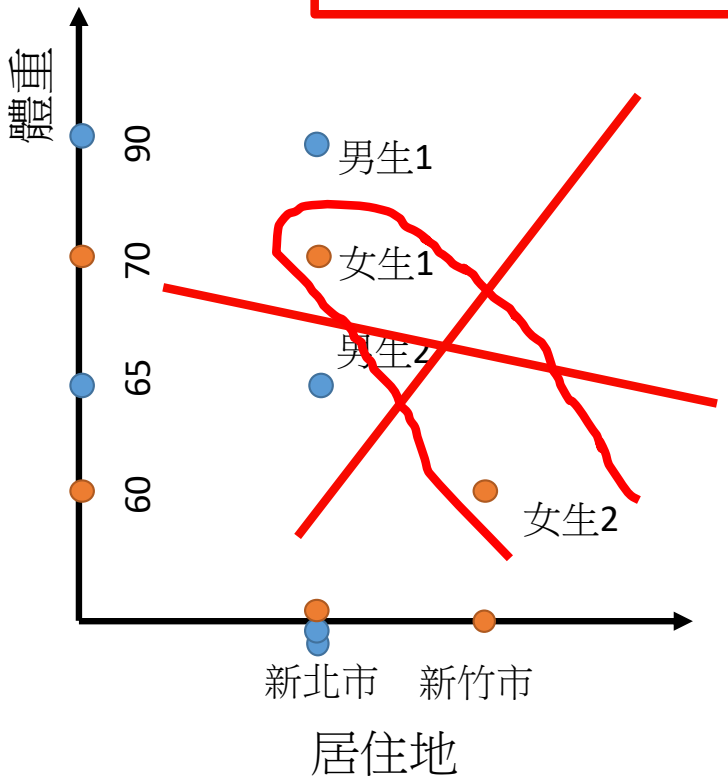
# Data

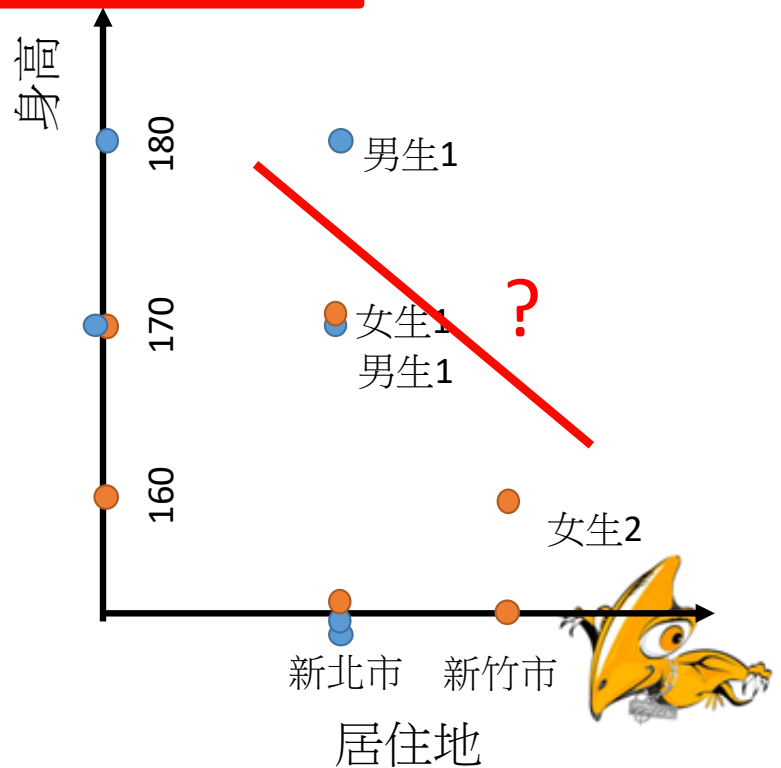
| | 居住地 | 身高 | 體重 |
|---|---|---|---|
| 男生1 | 新北市 | 180 | 90 |
| 女生1 | 新北市 | 170 | 70 |
| 男生2 | 新北市 | 170 | 65 |
| 女生2 | 新竹市 | 160 | 60 |

# Data

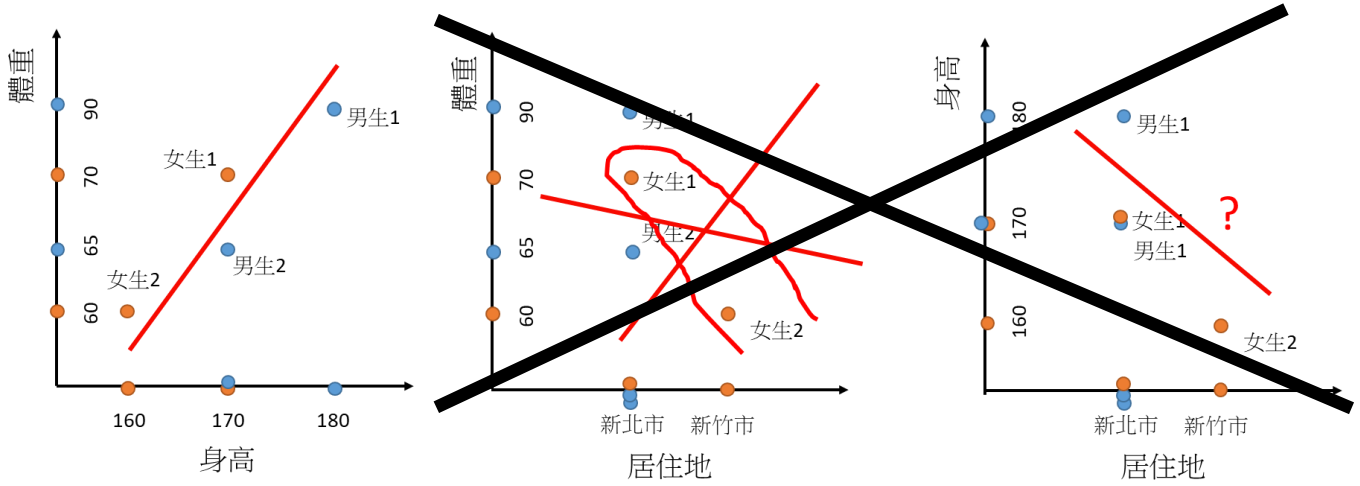
| | 居住地 | 身高 | 體重 |
|---|---|---|---|
| 男生1 | 新北市 | 180 | 90 |
| 女生1 | 新北市 | 170 | 70 |
| 男生2 | 新北市 | 170 | 65 |
| 女生2 | 新竹市 | 160 | 60 |



# Data

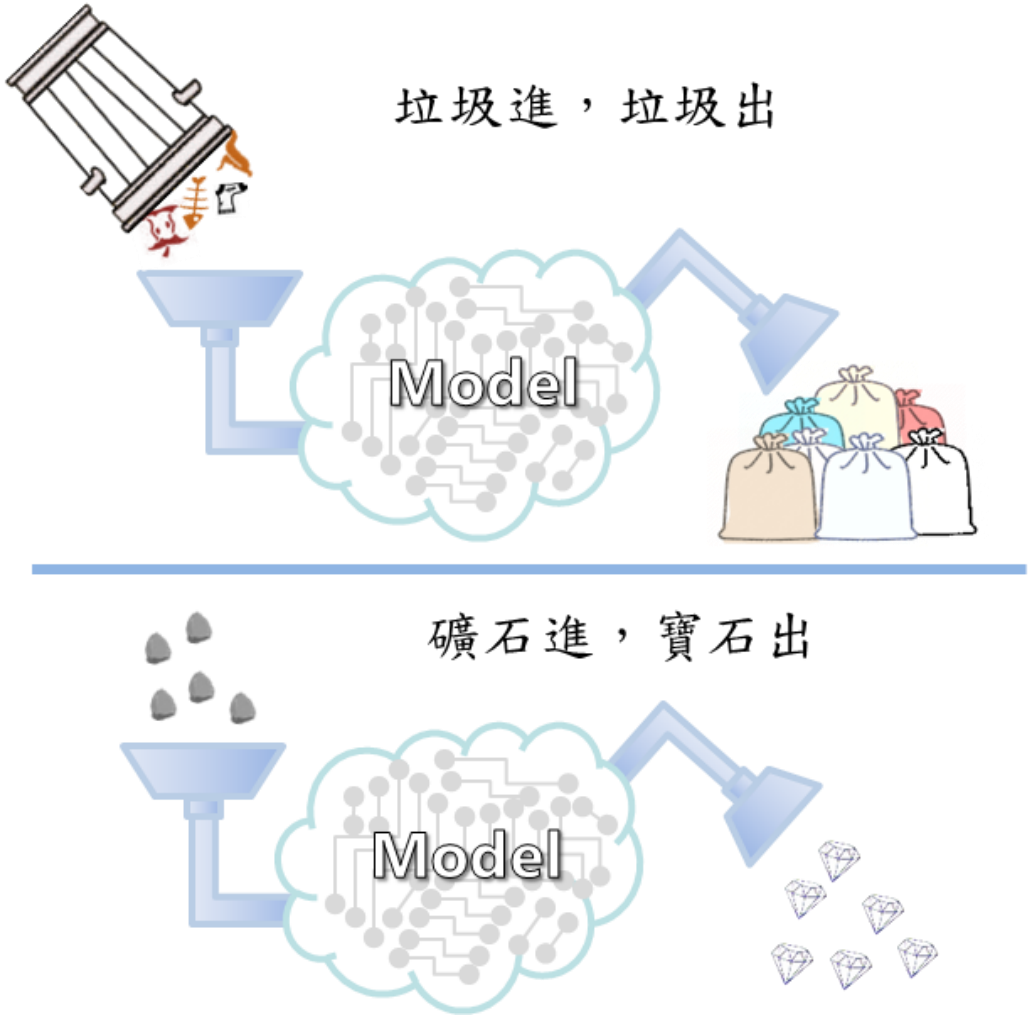



Data processing is also a science.

# Data

"MOST IMPORTANT" is DATA.

"**Garbage in, garbage out.**"



# Data

- How many data do I need in my projects?

**Question is what is your population?**

- For instance,

Fingerprint anti-spoof

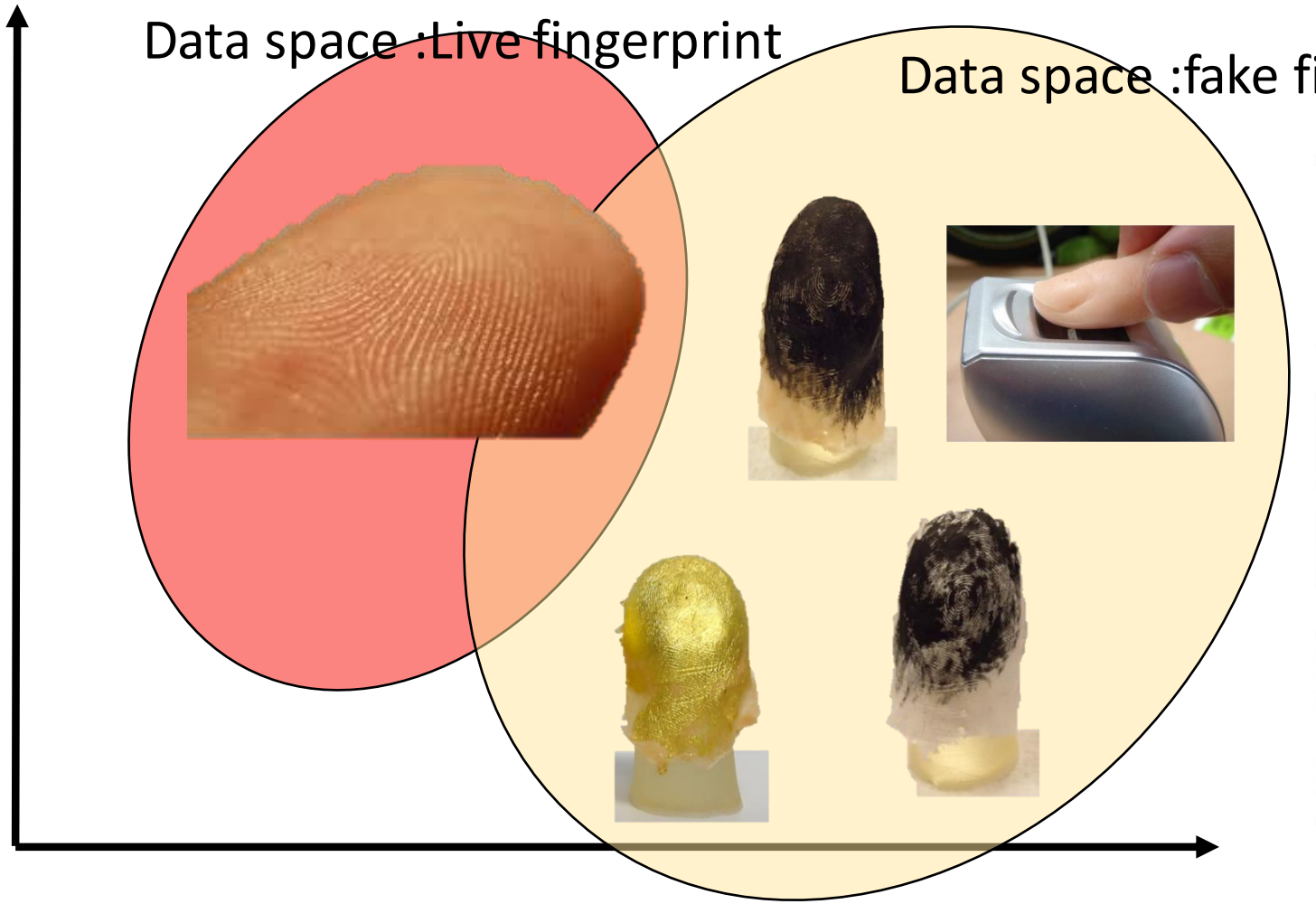
Figure 1: Conductive artifact – latex face paint with edible gold leaf coating

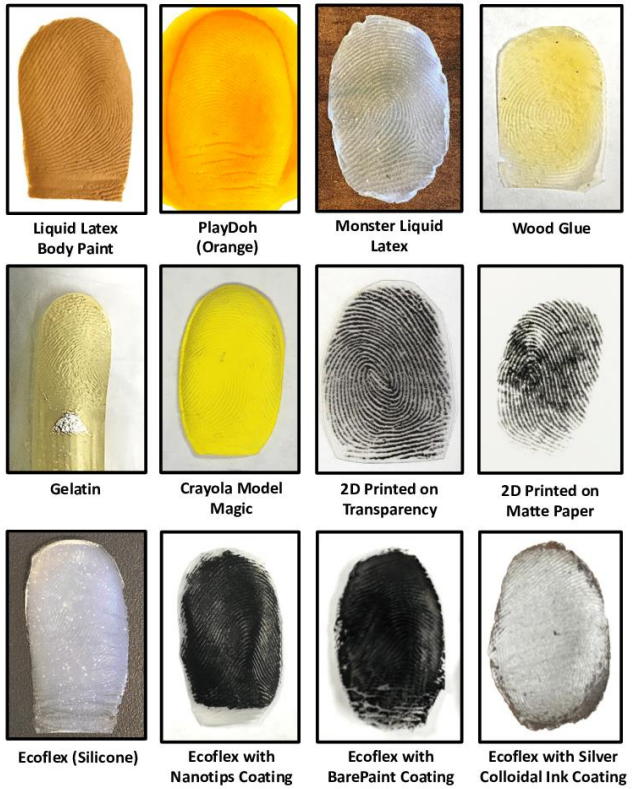
Figure 2: Conductive artifact – latex face paint with Bare Conductive Paint coating

Figure 3: Conductive artifact – acrylic paint with Bare Conductive Paint coating


# Data

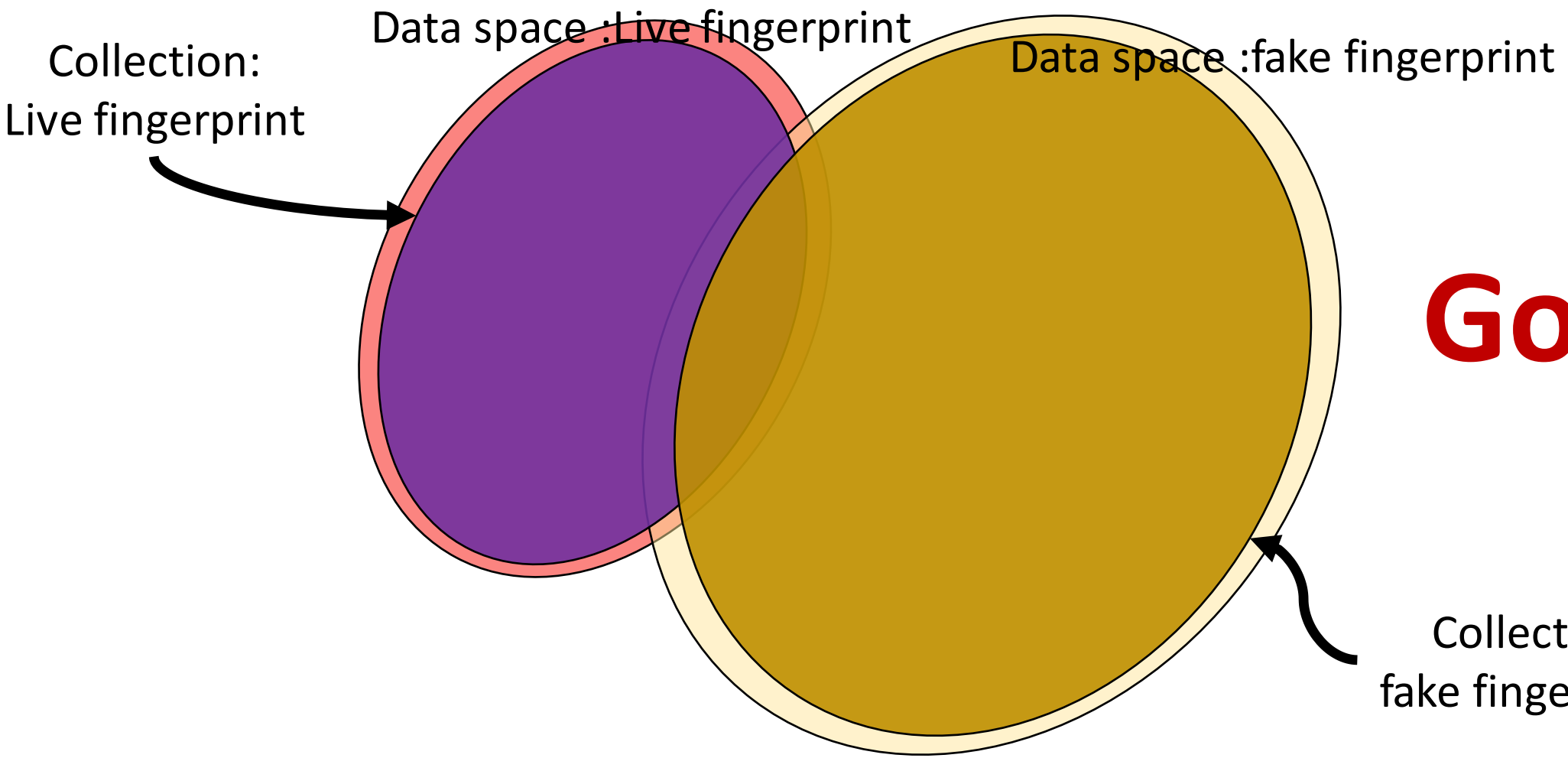
## Example: Fingerprint anti-spoof



# Data

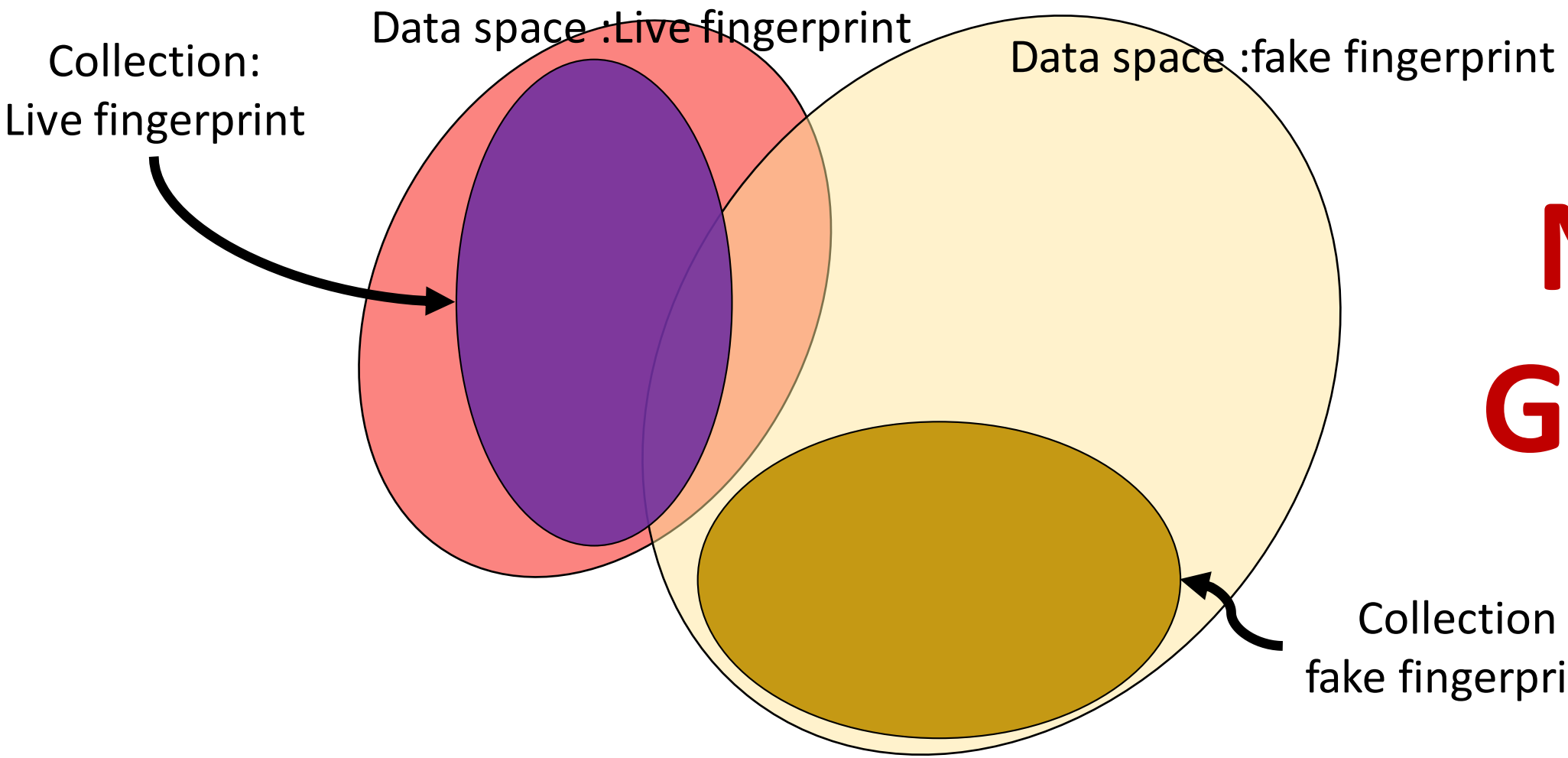
## Example: Fingerprint anti-spoof




**Good.**

# Data

## Example: Fingerprint anti-spoof

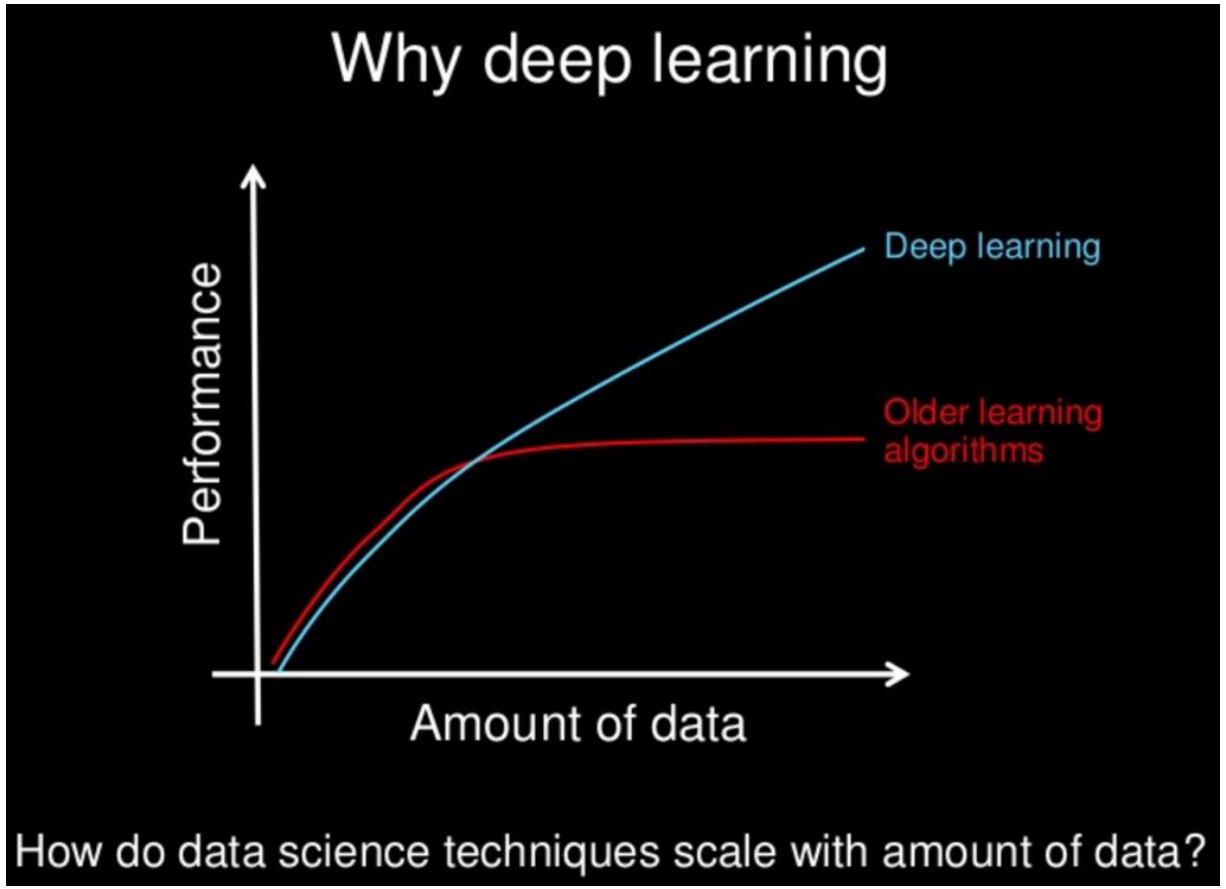



**Not Good.**

# Data

上述屬於Data的品質

Data的量?




Source: Andrew Ng,

# Data

- How many data do I need in my projects?

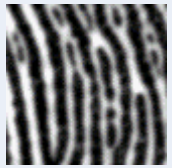
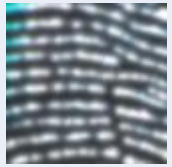
**Question is what is your population?**

- How to know the population? From sampling.

1. CLT (中央極限定理): randomly sampling as large as possible.
2. Metadata-based sampling: randomly sampling with different scenarios.
3. Manual curation: sample data based on metadata, and manually select the most "useful" data.

# Data

Metadata

| Images | Material | Recording device |
|---|---|---|
|  | Real | Capacitive |
|  | Gelatin | Capacitive |
|  | Real | Optical |
|  | Silcone | Optical |

# Data

- How many data do I need in my projects?

  **Question is what is your population?**

- How to know the population? From sampling.
- Can you make sure **the data pool** approximating to population?

ANS: No, we can't.

# Data

- Can you make sure **the data pool** approximating to population?

ANS: No, we can't.

**Based on metadata (scenarios definition), sampling data the more, the better.**

The collection dataset will be randomly separated to

1. Training set
2. Testing set
3. Validation set (Deep learning)

# Data

Based on metadata (scenarios definition), sampling data the more, the better.

IF we got a data pool of >1 million, which possibly contains noisy or useless information, a smart selection is required. (**Real Application**).

Keep studying and researching.

# Data (collection)

- 學術界:

Research-based

1. Open database (ImageNet, MSCOCO, VOC, Cityscape)
2. Data Collection by researchers (limits)

- 工業界:

Product-based

1. Open database (pre-trained)
2. Data collection in candidate device. (Device difference might be tested)
3. All sceneries must be considered.

# How to start a machine/deep learning application

- 1. Tasks/Applications definition.
- 2. Data (collection, labeling)
- 3. Learning Model
- 4. Evaluation model

# Learning Model

**Machine Learning algorithms** vs **Deep Learning algorithms**

Is necessary everything deep learning?

結構資料: 傳統的機器學習(SVM, Linear regression, LDA等)

非結構資料：深度學習。

# Learning Model

## 怎麼選模型?

用簡單的方法先做

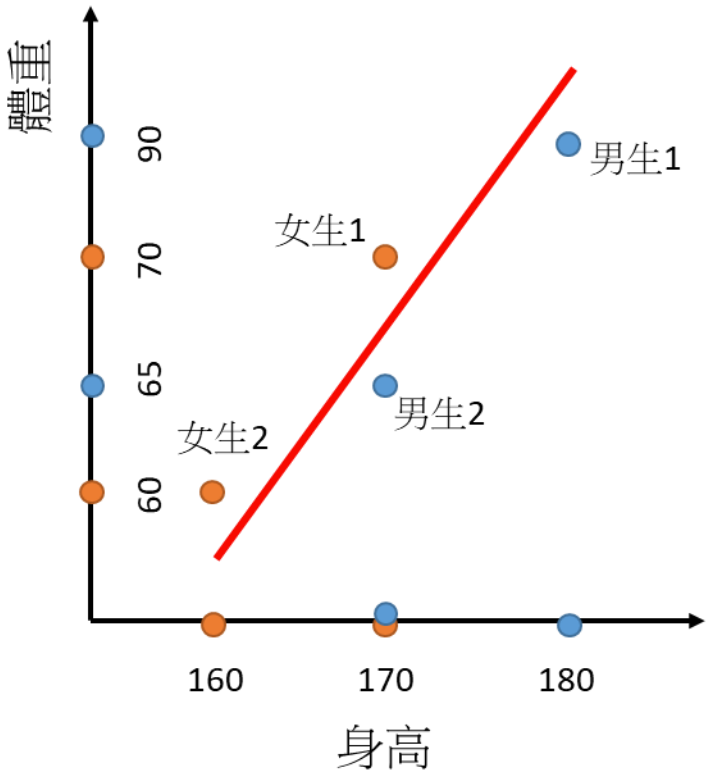
原因: 簡單的測試做得好，用複雜的只會更好，所以用最差的先做快速測試。

基線(簡單的方法)，後續會用不同常用的模型來進行訓練比較，然後挑一個模型。

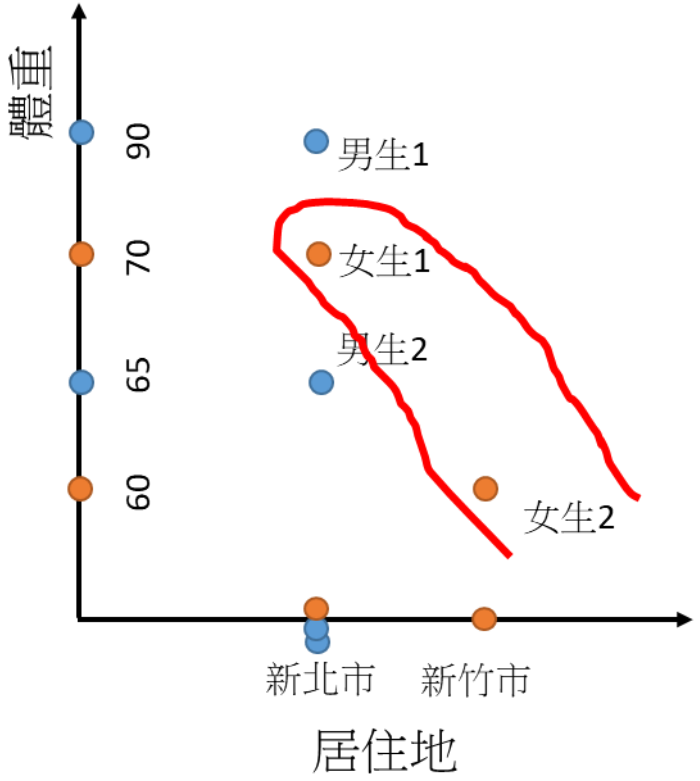
在機器學習上沒有一個方法是**最好的且不同領域都通用**的。

# Learning Model

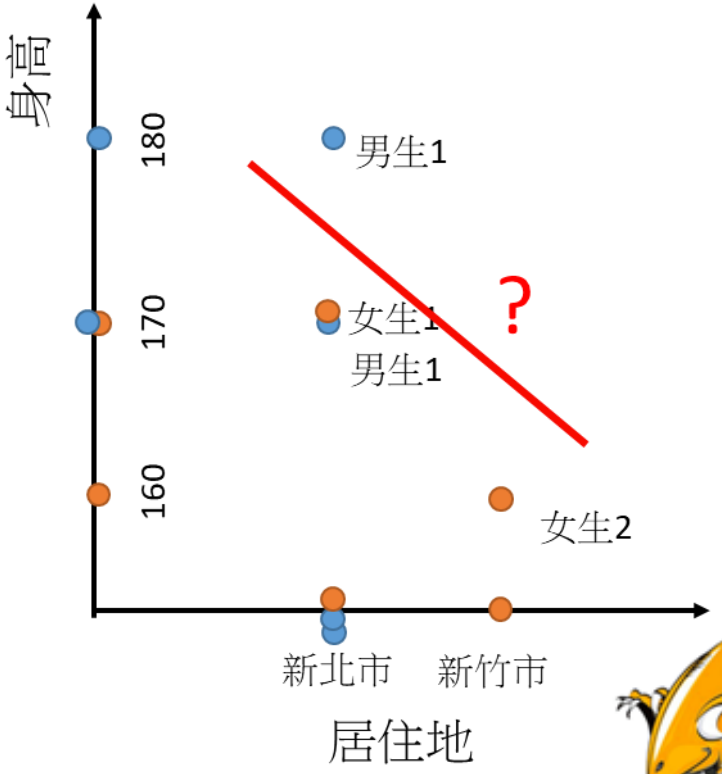
特徵選得好，**簡單的線性模型**就可以分得很好，不需要複雜的模型




特徵選得不太好，需要複雜一點的**非線性模型**




特徵選得非常不好，複雜一點的非線性模型也不能分類的好。



# Learning Model

## 怎麼選模型?

| 結構化數據: | 簡單的模型 | 複雜的模型 |
|---|---|---|
| Regression | Linear regression | SVR, GDBT, XGBoost |
| Classification | Linear Discriminant Analysis / logistic regression | SVM, GDBT, XGBoost |

| 非結構化數據: | 簡單的模型 | 複雜的模型 |
|---|---|---|
| | Backbone Network: MobileNet<br>Object detection: SSD<br>Segmentation: Unet | Backbone Network: Inceptionv4<br>Object detection:YOLOv4<br>Segmentation: HRNet |

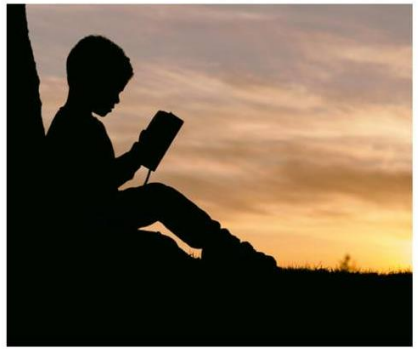
# Learning Model

## 怎麼選模型?

**1. 天縱英才**: 題目一來就知道用什麼算法。

**2. 後天努力**: 閱讀學習，大腦充滿一堆知識。

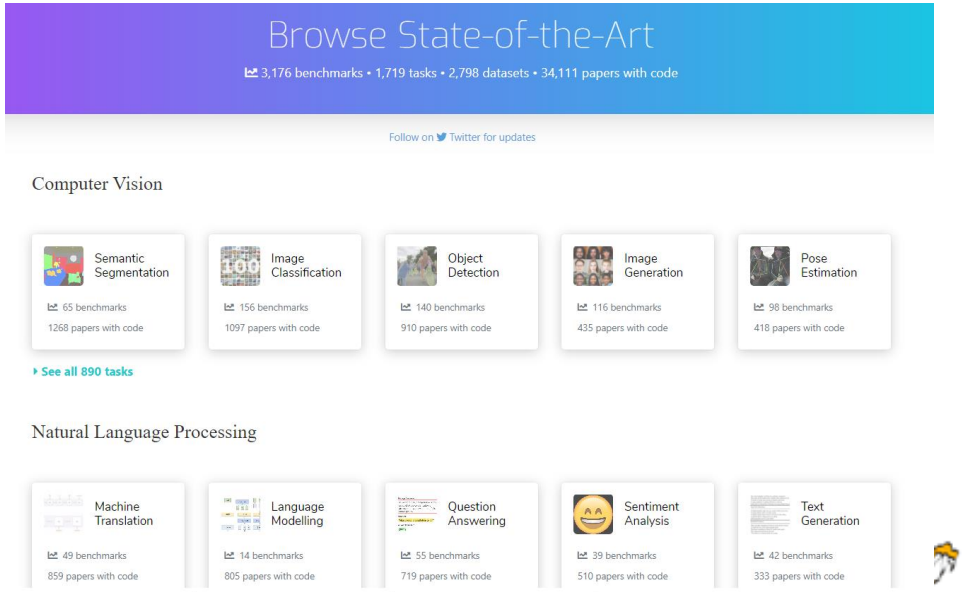
**3. Reference**: 別人怎麼做你就怎麼做。




主動學習新知 Human Learning → 了解 AI 應用及其影響 Machine Learning → 學習如何學習 Human Learning+

圖片來源:李孟

Paper with code



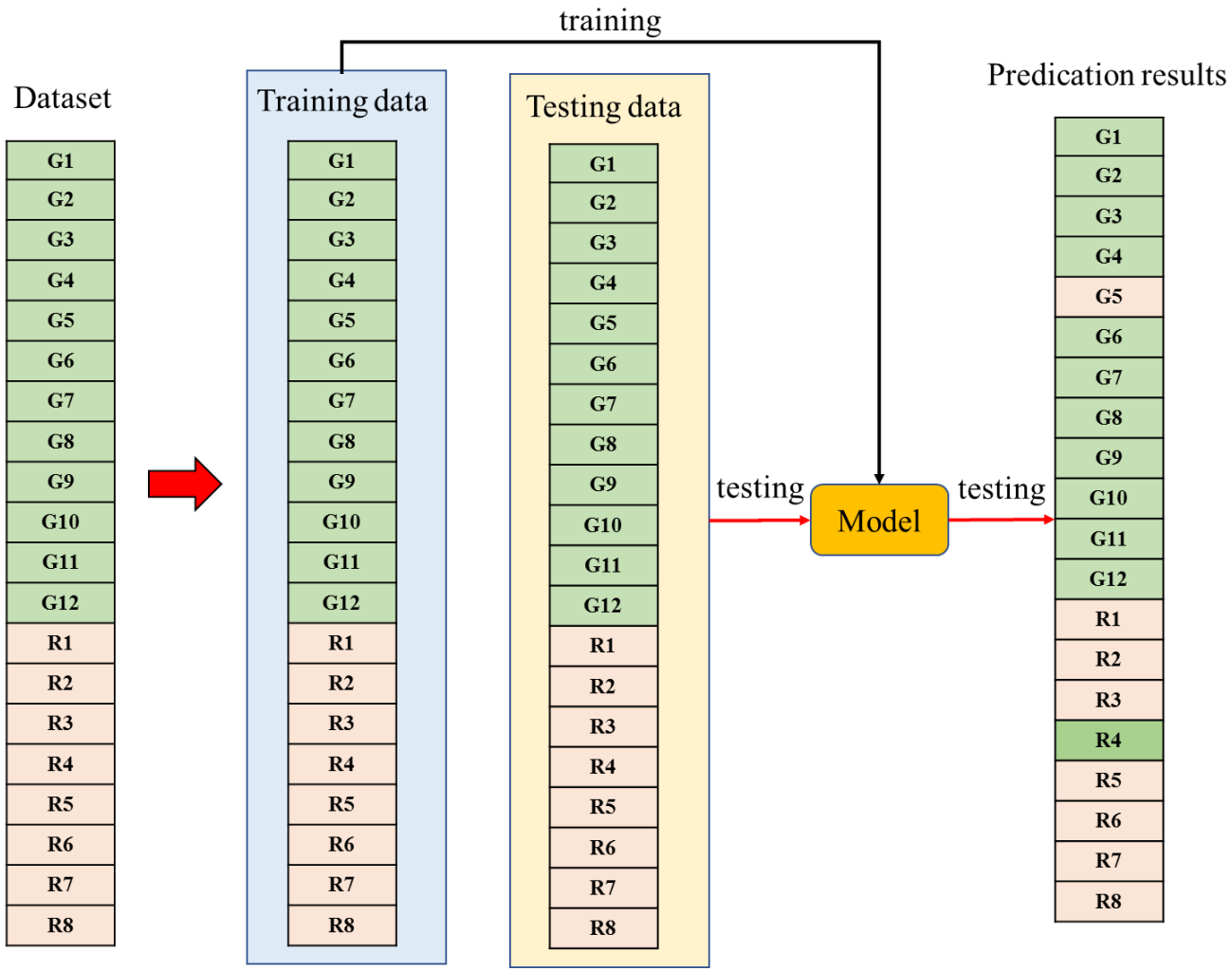
# How to start a machine/deep learning application

- 1. Tasks/Applications definition.
- 2. Data (collection, labeling)
- 3. Learning Model
- 4. Evaluation model

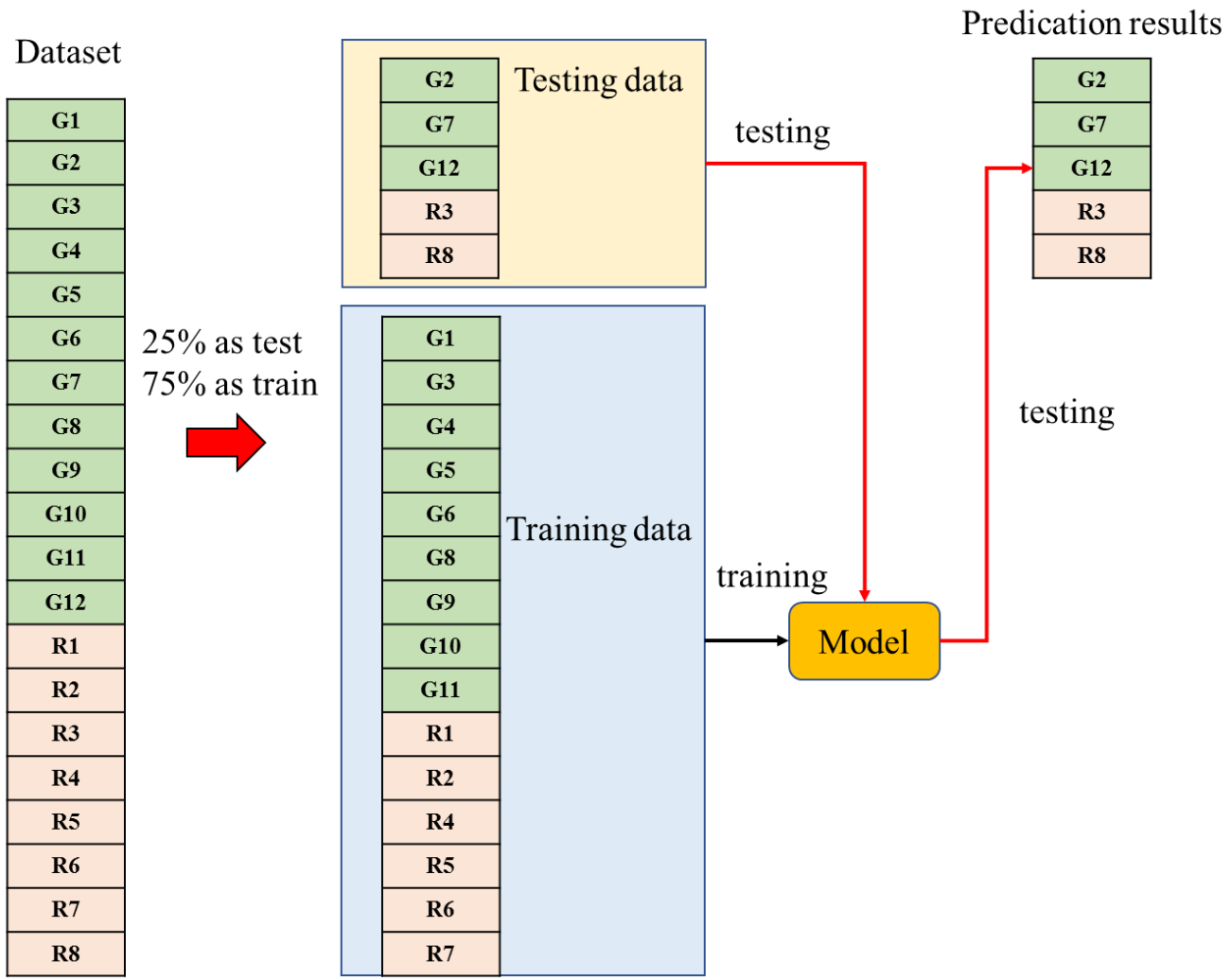
# Evaluation method
# 交叉驗證(Cross-validation, CV)

- 交叉驗證在機器學習上通常是用來驗證「你設計出來模型」的好壞。
- 1. Resubstitution
- 2. Holdout CV : Deep learning幾乎都是這種。
- 3. Leave-one-out CV
- 4. K-fold CV

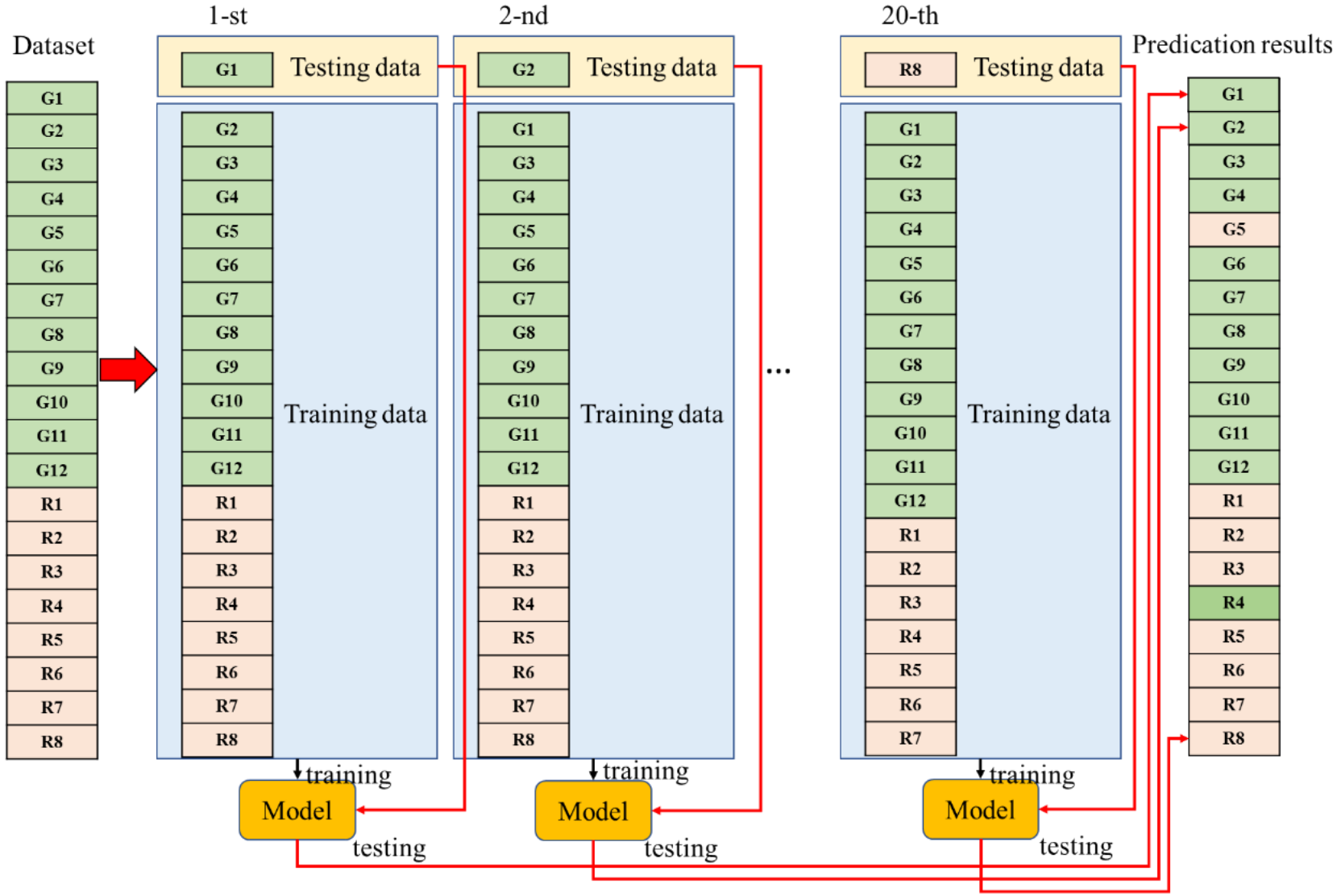
# 交叉驗證 - Resubstitution



# 交叉驗證 - Holdout CV

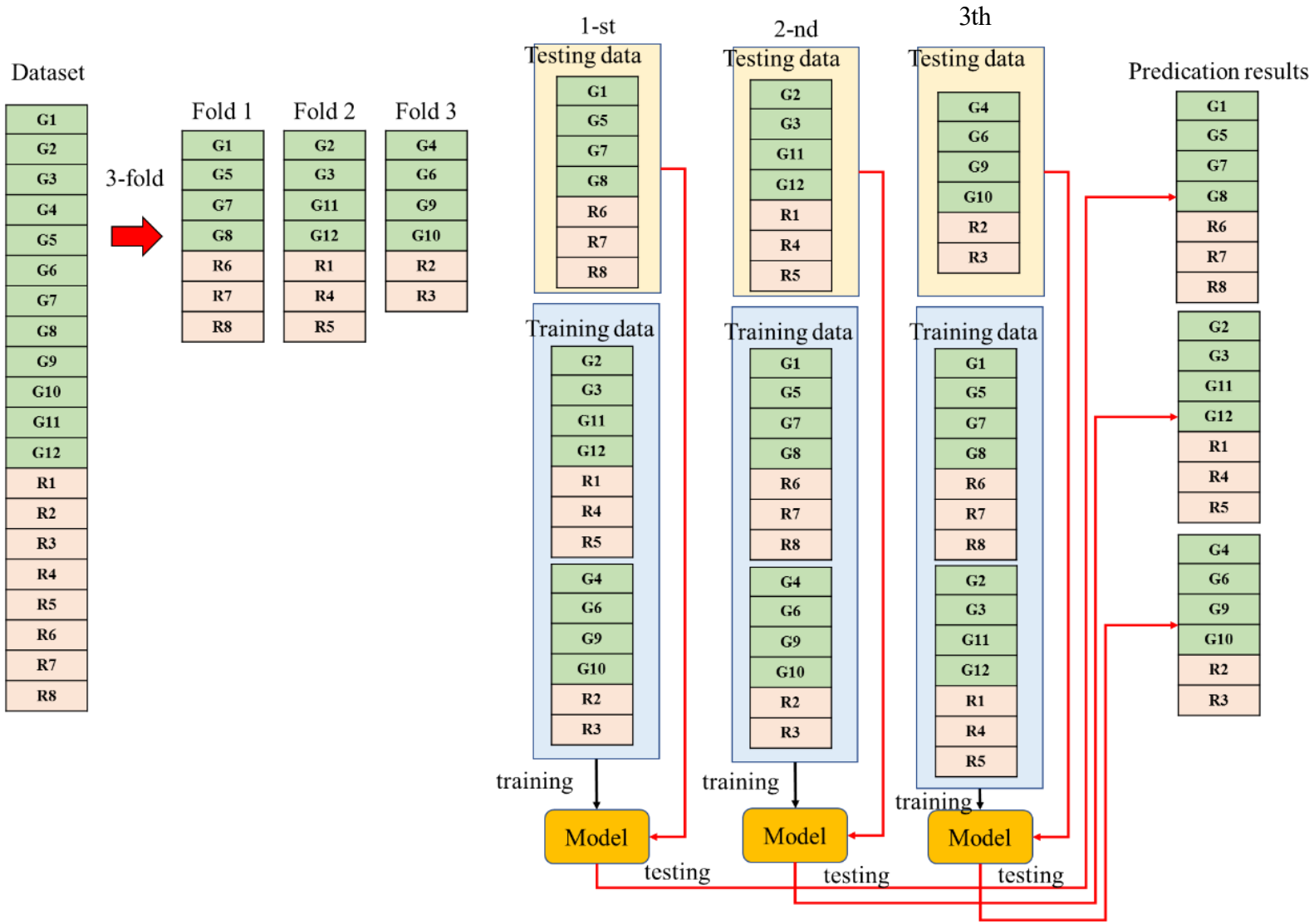


# 交叉驗證 - Leave-one-out CV



# 交叉驗證 - K-fold CV

## 3-fold CV



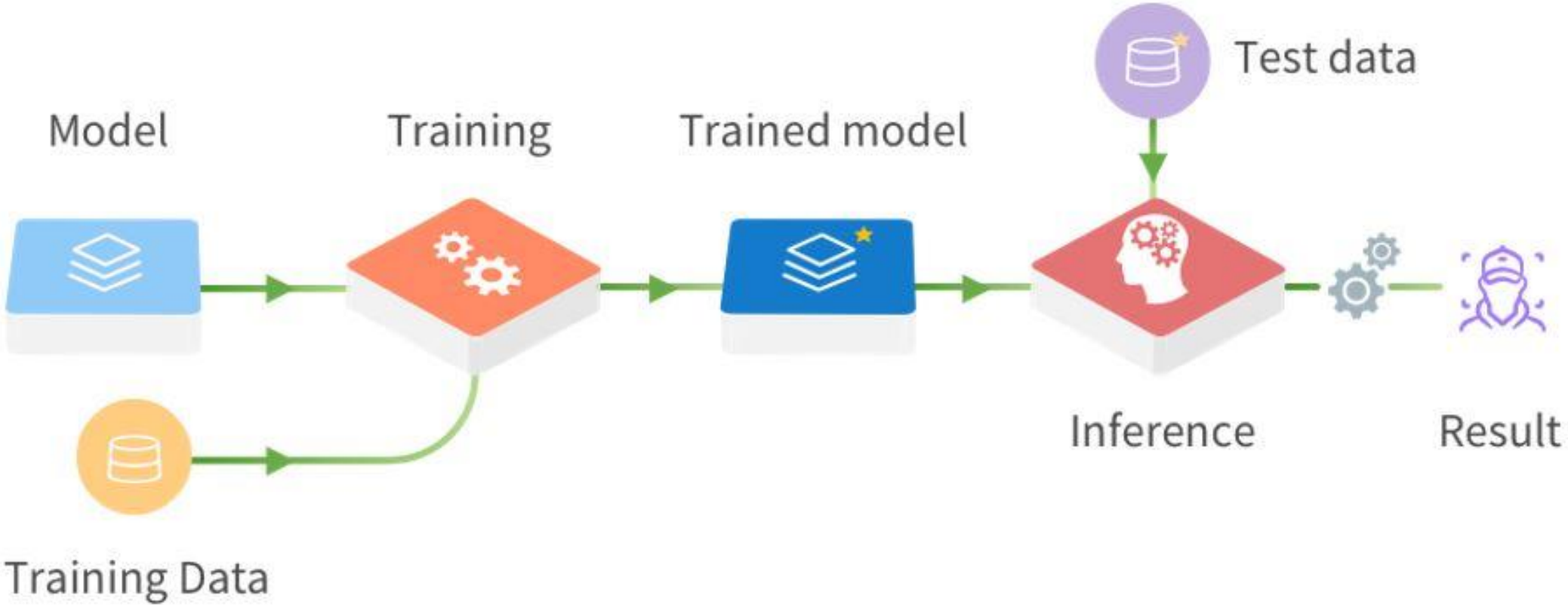
# Evaluation model

The collection dataset will be randomly separated to

1. Training set (抽樣盡可能逼進母體)
2. Testing set (盡可能跟實際會應用的資料一樣)
3. Validation set (Deep learning)

如果RD驗證的測試集準確度都沒達標，開發的演算法就不需要到測試部門進行測試。

# Evaluation model



# Evaluation model

**如果測試結果都不好，複雜的模型也做不出來。**

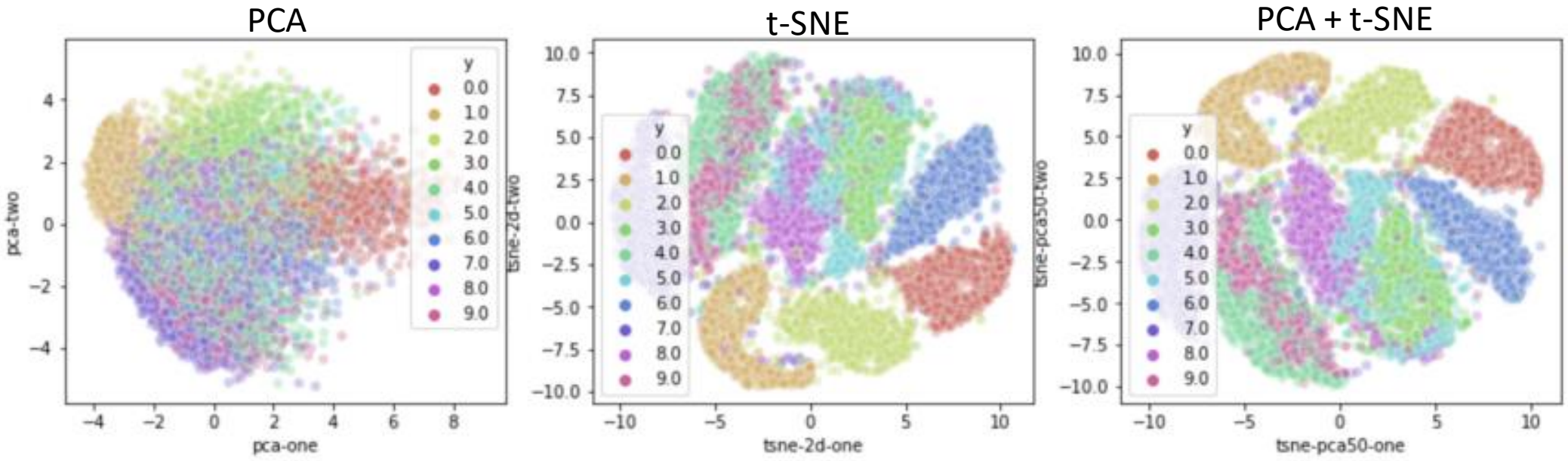
請回過頭來，

1. 檢視目的問題和Sensor是否是合理的。(不應該發生，在開案前就應該先釐清)
2. 檢視收集的數據。(最常發生: class-imbalance)

- 如果是結構化數據，請採用Visualization技術看訓練和測試資料的分布情形。
- 非結構數據，請分析資料判斷的結果，補足數據。

# Evaluation model

結構化數據，請採用Visualization技術看訓練和測試資料的分布情形。
結構數據是多維度理論上無法用視覺呈現，因此可以採用PCA或是t-SNE。




MNIST flatten後進行資料壓縮

https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b

# Evaluation model

非結構數據，請分析資料判斷的結果，補足數據。

**檢視收集的數據：最常發生class-imbalance**

假設 數據量：正樣本資料9999萬筆，負樣本1萬筆。

深度學習進行batch learning。

一個batch(1000筆): 正樣本有999筆，負樣本1筆。

假設正樣本loss梯度都是0.01，負樣本loss梯度為1，平均後梯度為0.010099。

模型在update的時候，負樣本的梯度會被正樣本的平均掉，怎麼學都學不到負樣本。

# Q & A