



# Task 4

## Synthetic Data Vault

By 白宸宇

# Overview

01

## Single Table

使用Tabular Model,  
合成單一筆資料集中的欄位,

對應非潛在客戶的資料  
non\_potential\_customers

02

## Multi Table

使用Relational Model,  
合成跨資料集的多重資料。

對應潛在客戶的資料  
potential\_customers和  
ads.csv有相同的user\_id作為  
key

03

## Sequential

使用Sequential及Time series  
Model,  
合成時間依賴性或序列性的資  
料。

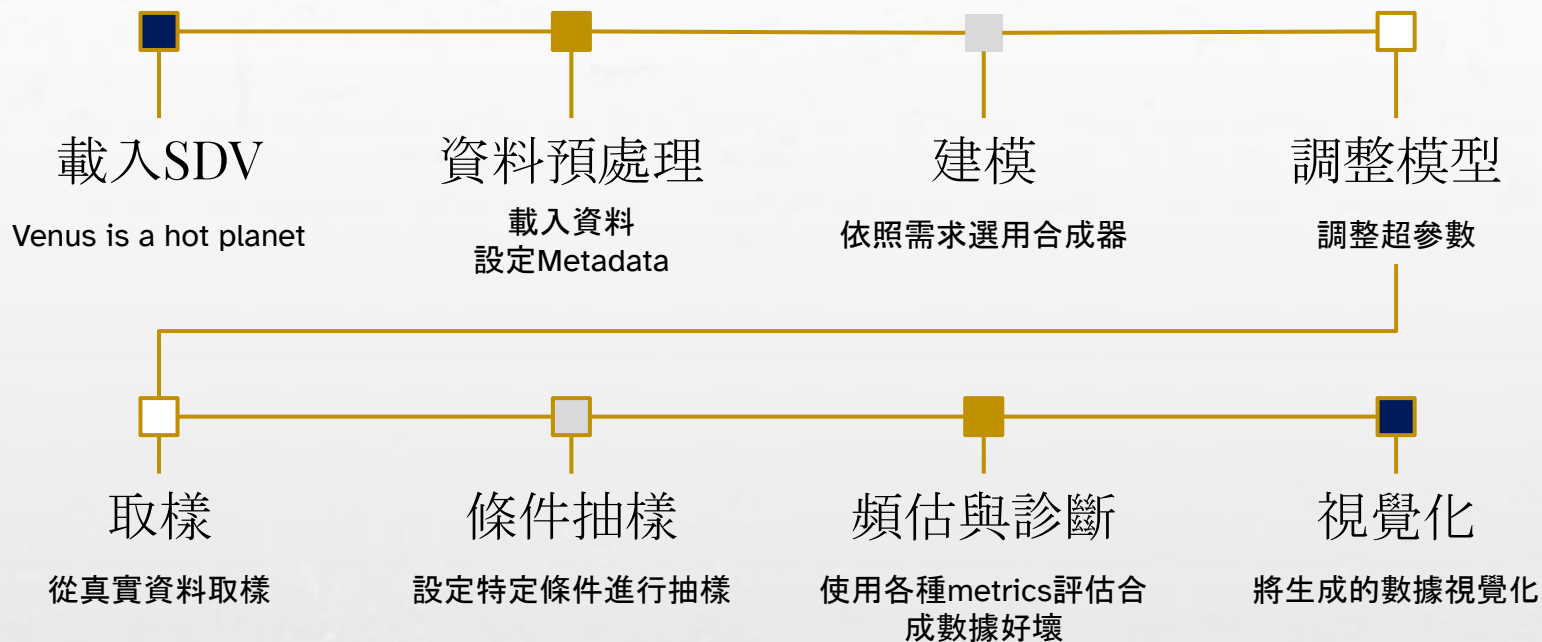
對應兩資料集中的時間戳資料  
e\_et、pt\_d

# The SDV ecosystem

	Type	Description	Usage
<b>Copulas</b>	Tabular	使用 copula 函數來描述多變量數據之間的相依性，並根據這些相依性生成新的數據。	生成表格內較簡單的合成數據
<b>CTGAN</b>	Tabular	基於生成對抗網絡(GAN)的模型，專門用於生成高保真度的合成表格數據。	生成表格內較複雜的合成數據
<b>DeepEcho</b>	Time Series	結合了經典的統計模型和深度學習方法，能夠生成遵循時間依賴性的數據	生成時間戳的合成數據
<b>RDT</b>	Transforms	將原始資料轉換為數值資料以便用於資料科學專案，需要時將其轉回原始格式	可將涉及隱私問題的PII匿名化

另有提供SDGym來做Benchmarking、SDMetrics來評估績效

# 流程圖(Single Table)



# Data Preparation



# Modeling

## Gaussian Copula

- 使用高斯分布和Copula函數來捕捉變數間的依賴性，分離邊際分佈和依賴結構。。
- 適合數值和連續變數的合成。
- Pros:能夠處理複雜的相關結構。

## CTGAN

- 基於生成對抗網絡(GAN)，專為表格數據設計。
- 擅長處理複雜分佈和高稀疏性數據。
- Pros:生成高保真度的合成數據。

## TVAE

- 基於變分自動編碼器，適合生成連續和分類數據的混合數據集。
- 採用概率方法來表示隱變數並重建數據。
- Pros:處理非線性關係和混合數據類型。

# Sampling

## Direct sampling

直接從合成模型生成的數據分佈中隨機抽樣，不考慮特定條件。

適用於需要生成與整體數據相似的數據集。

## Conditional sampling

根據指定的條件或特徵(如特定變數的值)生成數據。

能夠更精細地控制合成數據的屬性，使其符合特定需求或分析場景。例如：設定年齡分佈的比例

EX:可以設定只要生成Task\_count大於一定數量的數據，亦或生成特定時間戳的資料。

# Evaluation

01

## Diagnostic

執行基本的**數據格式**和有效性檢查，確保合成數據是有效的。

EX:

有效性: Primary Key唯一性、連續值的範圍、離散值的類別。

數據結構: 真實數據和合成數據的欄位名稱一致性。

02

## Data Quality

評估合成數據與真實數據在統計上的相似性。試圖理解合成數據在多大程度上反映了真實數據的屬性。

主要測量數據的**統計特徵**，如分布和相關性。

03

## Visualization

進行視覺化有助於更直觀地識別數據**分佈和趨勢**之間的差異。



# Part 3

```
# Step 2: Train the synthesizer
synthesizer.fit(non_potential_customers_2)

# Step 3: Generate synthetic data
synthetic_data = synthesizer.sample(num_rows=100)
```

✓ 3m 37.2s

```
~/opt/anaconda3/lib/python3.9/site-packages/sdv/lite/single_table.py in <module>
```

```
8 import cloudpickle
```

```
9
```

```
----> 10 from sdv.single_table import GaussianCopulaSynthesizer
```

```
11
```

```
12 LOGGER = logging.getLogger(__name__)
```

```
...
```

```
3
```

```
4 from ..common import _linalg
```

```
5 from .._internal import get_xp
```

```
ImportError: cannot import name '__all__' from 'numpy.linalg' (/Users/castle/opt/anaconda3/lib/python3.9/site-packages/numpy/linalg/__init__.py)
```

Output is truncated. View as a [scrollable element](#) or open in a [text editor](#). Adjust cell output [settings](#)..



**Thanks for watching**

