



# U8-Lab2. Datos Sintéticos

# ÍNDICE

---

# LAB2. Datos Sintéticos



1. Objetivos
2. Motivación de los datos sintéticos
3. Ciclo de vida de la creación de datos sintéticos
4. Supuesto práctico
  - a. Datos reales
  - b. Herramienta
  - c. Google Colab

---

# LAB2. 1. OBJETIVOS

# LAB2. 1. Objetivos



## Objetivos:

- Tomar consciencia de las principales motivaciones para usar datos sintéticos
- Aprender a utilizar una herramienta para la creación de datos sintéticos: SDV (Synthetic Data Vault),

---

## 2.- Motivación de uso de datos sintéticos

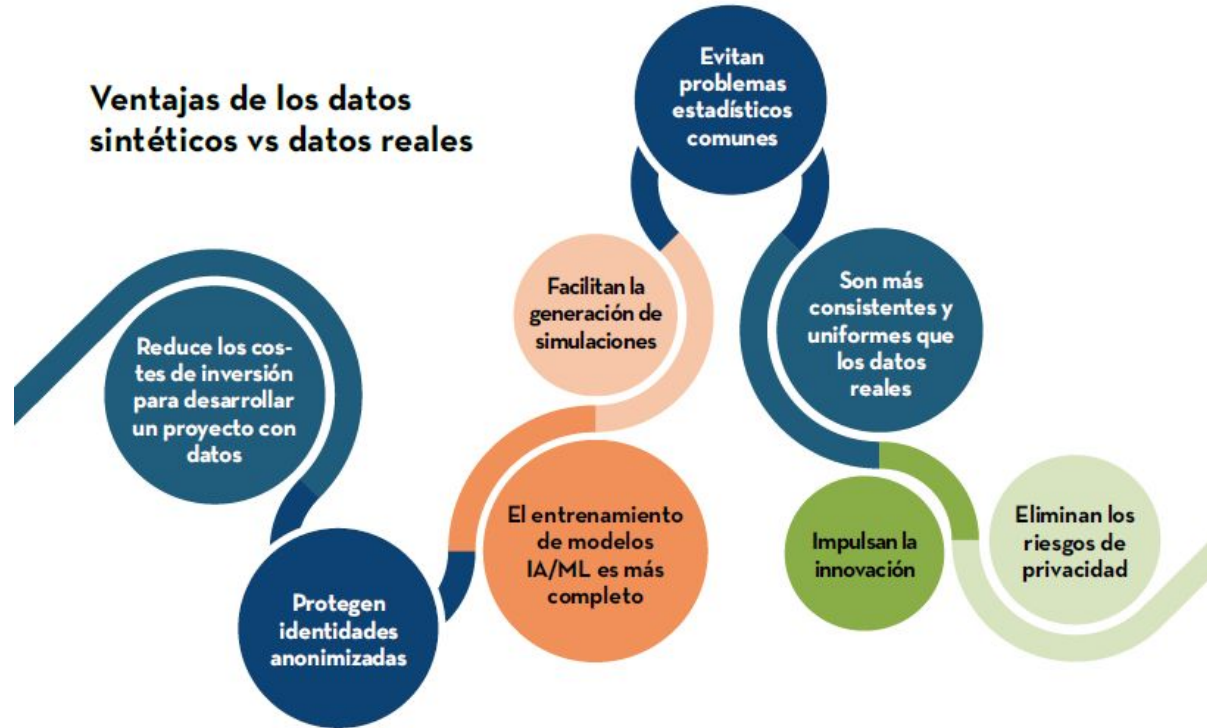
# LAB2. 2. Motivación del uso de datos sintéticos



1. Superar restricciones regulatorias
2. Preservación de la privacidad
3. Resistencia a la reidentificación
4. Facilitar la innovación y monetización
5. Agilizar la simulación
6. Lograr una mayor consistencia
7. Facilitar el entrenamiento de modelos y permitir una manipulación sencilla
8. Viabilizar proyectos incipientes y aumentar la rentabilidad.

[Vídeo del informe](#)

# LAB2. 2. Motivación del uso de datos sintéticos





---

## LAB2. 3.- Flujo de trabajo para la creación de datos sintéticos

## LAB2. 2. 3. Flujo de trabajo para creación de Datos Sintéticos



---

## LAB2. 4.- Supuesto práctico

link [a GoogleColab](#)

# LAB2. 4.a Datos Reales



Nuestro conjunto de datos corresponde a los clientes alojados en un hotel

guest_email	has_rewards	room_type	amenities_fee	checkin_date	checkout_date	room_rate	billing_address	credit_card_number
michaelsanders@shaw.net	False	BASIC	37.89	27 Dec 2020	29 Dec 2020	131.23	49380 Rivers Street\nSpencerville, AK 68265	4075084747483975747
randy49@brown.biz	False	BASIC	24.37	30 Dec 2020	02 Jan 2021	114.43	88394 Boyle Meadows\nConleyberg, TN 22063	180072822063468
webermelissa@neal.com	True	DELUXE	0.00	17 Sep 2020	18 Sep 2020	368.33	0323 Lisa Station Apt. 208\nPort Thomas, LA 82585	38983476971380
gsims@terry.com	False	BASIC	NaN	28 Dec 2020	31 Dec 2020	115.61	77 Massachusetts Ave\nCambridge, MA 02139	4969551998845740
misty33@smith.biz	False	BASIC	16.45	05 Apr 2020	NaN	122.41	1234 Corporate Drive\nBoston, MA 02116	3558512986488983

# LAB2. 4.a Datos Reales

Los metadatos asociados a nuestros datos son:

```
{
  "METADATA_SPEC_VERSION": "SINGLE_TABLE_V1",
  "primary_key": "guest_email",
  "columns": {
    "guest_email": {
      "sdtype": "email",
      "pii": true
    },
    "has_rewards": {
      "sdtype": "boolean"
    },
    "room_type": {
      "sdtype": "categorical"
    },
    "amenities_fee": {
      "sdtype": "numerical",
      "computer_representation": "Float"
    },
    "checkin_date": {
      "sdtype": "datetime",
      "datetime_format": "%d %b %Y"
    },
    "checkout_date": {
      "sdtype": "datetime",
      "datetime_format": "%d %b %Y"
    },
    "room_rate": {
      "sdtype": "numerical",
      "computer_representation": "Float"
    },
    "billing_address": {
      "sdtype": "address",
      "pii": true
    },
    "credit_card_number": {
      "sdtype": "credit_card_number",
      "pii": true
    }
  }
}
```


# LAB2. 4.a Datos Reales



Para asegurarnos de que los datos generados no contienen datos personales, todos los campos identificadores deben contener "pii":true. pii Personal Identifiable Information.

```
"credit_card_number": {  
    "sdtype": "credit_card_number",  
    "pii": true  
}
```

## LAB2. 4.b Herramienta

-  SDV(Synthetic Data Vault)
  - Proyecto Iniciado en el MIT en 2016
  - Librería de Python basada en aprendizaje automático para aprender patrones de sus datos reales y emularlos en datos sintéticos.
  - Métodos generativos para datos sintéticos(deep-learning)
  - Versión comercializada desde 2020 por DataCebo

## LAB2. 4.c Google Colab

1. Instalar la librería SDV

```
%pip install sdv
```

2. Cargar los datos

```
from sdv.datasets.demo import download_demo  
real_data, metadata = download_demo(  
    modality='single_table',  
    dataset_name='fake_hotel_guests'  
)
```



## LAB2. 4.c Google Colab

### 3. Creamos un sintetizador preajustado

```
from sdv.lite import SingleTablePreset  
  
synthesizer = SingleTablePreset(  
    metadata,  
    name='FAST_ML'  
)
```

## LAB2. 4.c Google Colab

- 
4. Entrenamos el sintetizador con los datos reales

```
synthesizer.fit(  
    data=real_data  
)
```

5. Utilizamos el modelo para crear nuevos datos

```
synthetic_data = synthesizer.sample(  
    num_rows=500  
)
```

## LAB2. 4.c Google Colab

### 6. Evaluamos los datos:

- a. Las columnas que hemos marcado con `pii=true` están completamente anonimizados. Los datos son falsos siguiendo el formato original
- b. El resto de las columnas replican las propiedades matemáticas de las columnas originales. Para evaluarlos se utiliza el módulo de evaluación. con una métrica para los valores categóricos y otra para los numéricos

```
from sdv.evaluation.single_table import evaluate_quality
quality_report = evaluate_quality(
    real_data,
    synthetic_data,
    metadata
)
```

## LAB2. 4.c Google Colab

### 6. Evaluamos los datos:

Column Shapes evalúa la similitud estadística entre los datos reales y sintéticos para columnas individuales de datos. A esto se le suele llamar distribución marginal de cada columna.

Column Pair Trends La similitud estadística entre los datos reales y sintéticos para pares de columnas. Esto a menudo se denomina correlación

## LAB2. 4.c Google Colab

7. Guardamos el sintetizador:

```
synthesizer.save('my_synthesizer.pkl')
```

8. Posteriormente podríamos volver a utilizarlo para generar más datos

```
synthesizer = SingleTablePreset.load('my_synthesizer.pkl')
```