

# **MODULO IV - MODELOS DE** **APRENDIZAJE SUPERVISADO**

DATA SCIENCE CON PYTHON

---

## ► Definición de Machine Learning

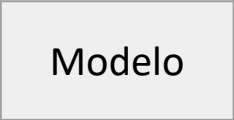
*"Es un programa de computador que aprende de la **experiencia E**, respecto a alguna **tarea T** y con medida de **rendimiento P**, si el desempeño sobre la tarea T, medido por P, mejora con la experiencia E.*

*Tom Mitchell - 1997*

Carnegie Mellon University



Pasado



Presente



Futuro

Etapa de aprendizaje

Experiencia



Modelo

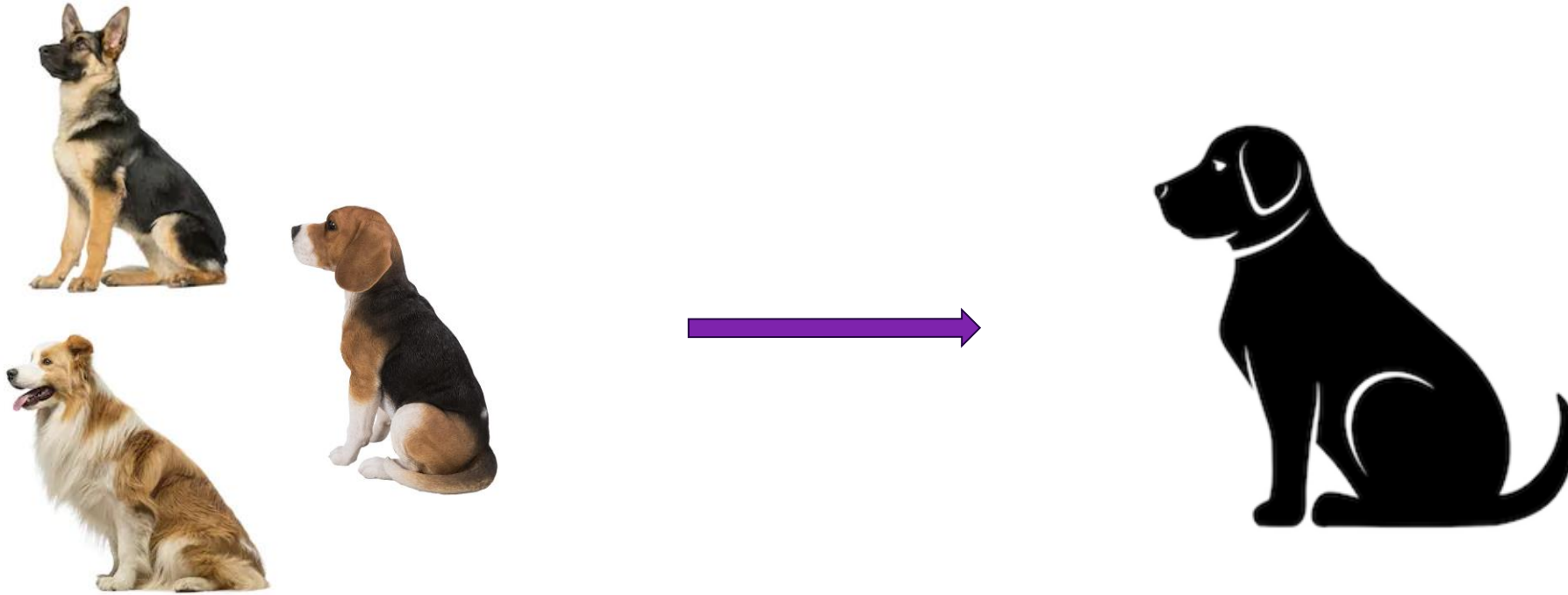
Etapa de predicción



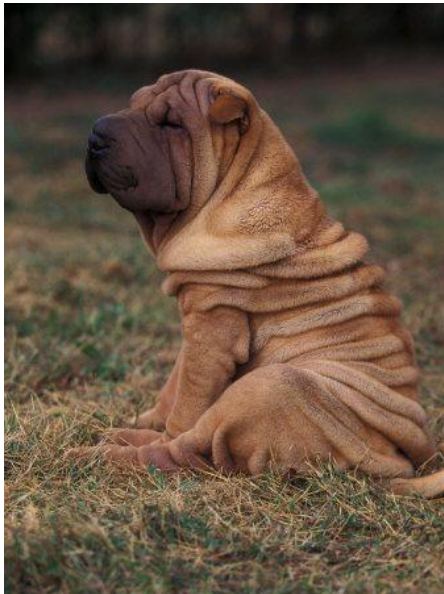
Predicciones

Un **modelo** es una representación de la realidad

Generamos nuestro **modelo** en la etapa de aprendizaje



Y luego en la etapa de predicción lo utilizamos contra nuevos ejemplos



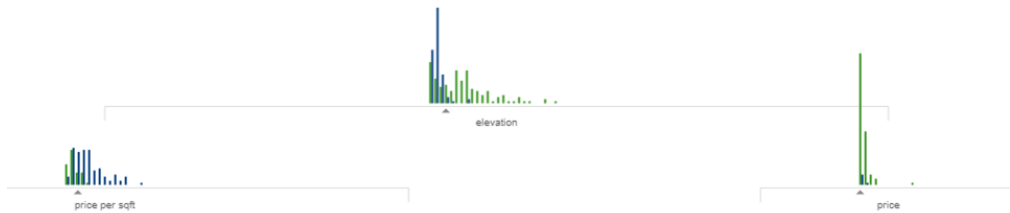
?



Si

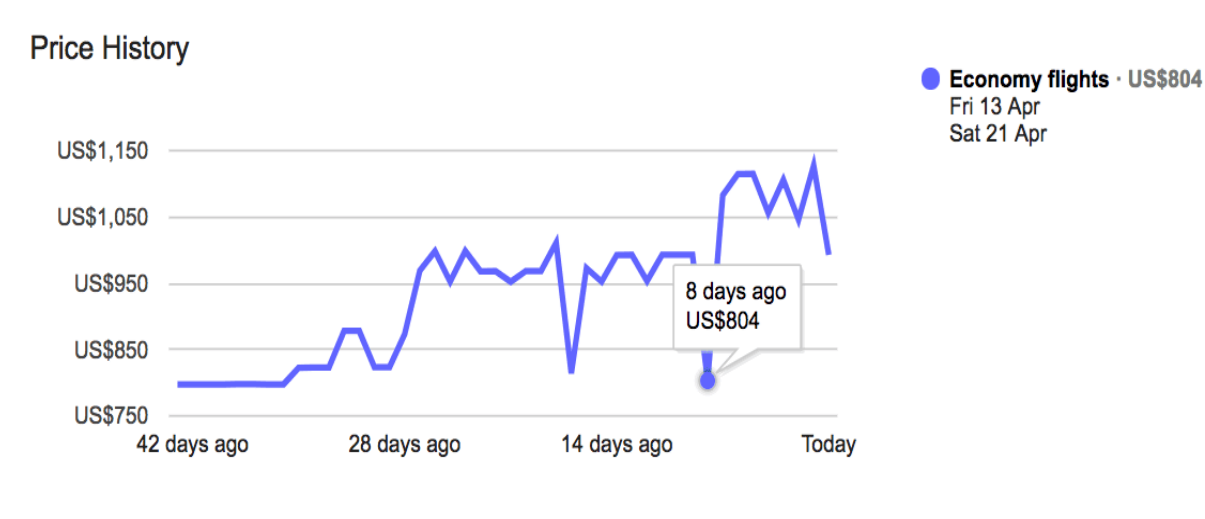
No

- ▶ Basado en el ejemplo de los perros, pero para hacerlo más divertido sumamos gatos.



- ▶ **Supervised Learning -Clasificación-**
  - ▶ Histórico etiquetado de perros y gatos.
  - ▶ Decidir, basado en el histórico, si una nueva imagen es un perro o un gato. **(Binaria)**
  - ▶ Si hay más animales en el histórico, clasificar por más clases. **(Multi-clase)**

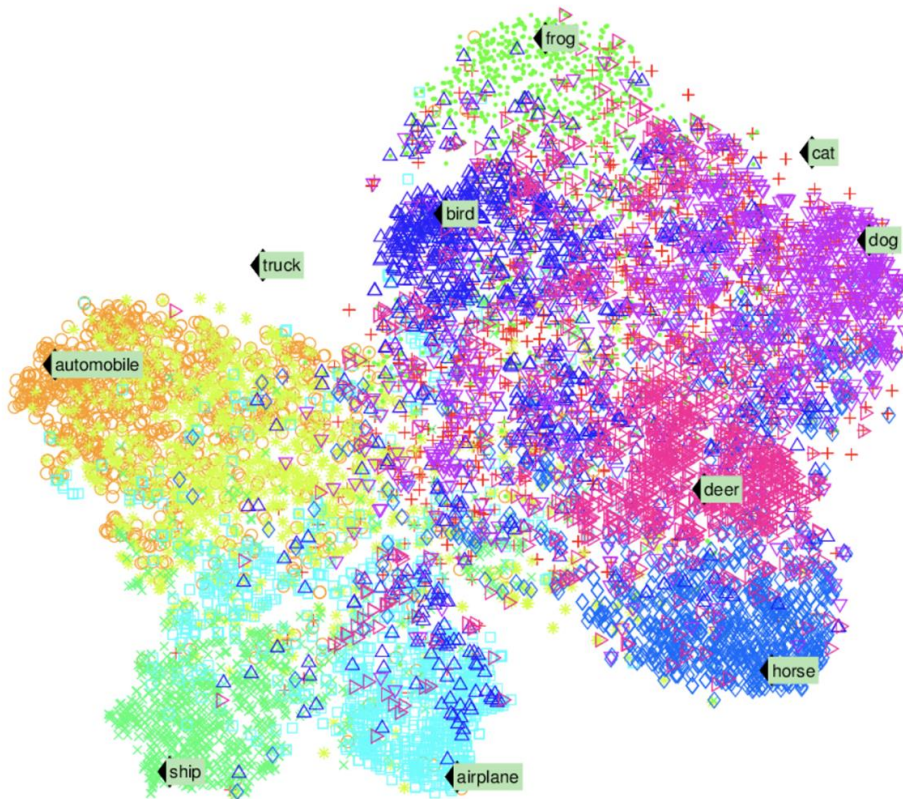
- ▶ Cuándo sería más barato viajar a Bali? Y cuándo compro el ticket?



- ▶ **Supervised Learning -Regresión-**
  - ▶ Histórico etiquetado de pasajes a Bali con su **precio**.
  - ▶ Decidir, basado en el histórico, cuándo es la mejor opción.

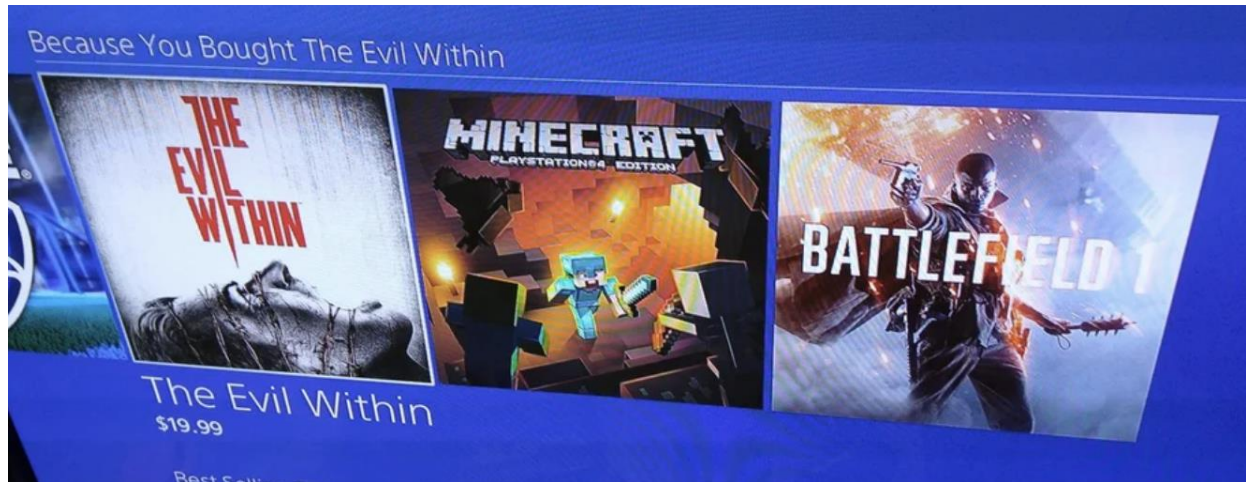


- Y si quiero una representación de los objetos que aparecen en el juego, pero hay demasiadas características para cada objeto?



- **Unsupervised Learning**  
**-Reducción de dimensionalidad-**
  - Representación 2D o 3D
  - Identifica los componentes principales de cada objeto.  
**(Simplificar sin perder tanta información)**

- ▶ Qué tal si queremos recomendarles juegos a gente afín a otra?



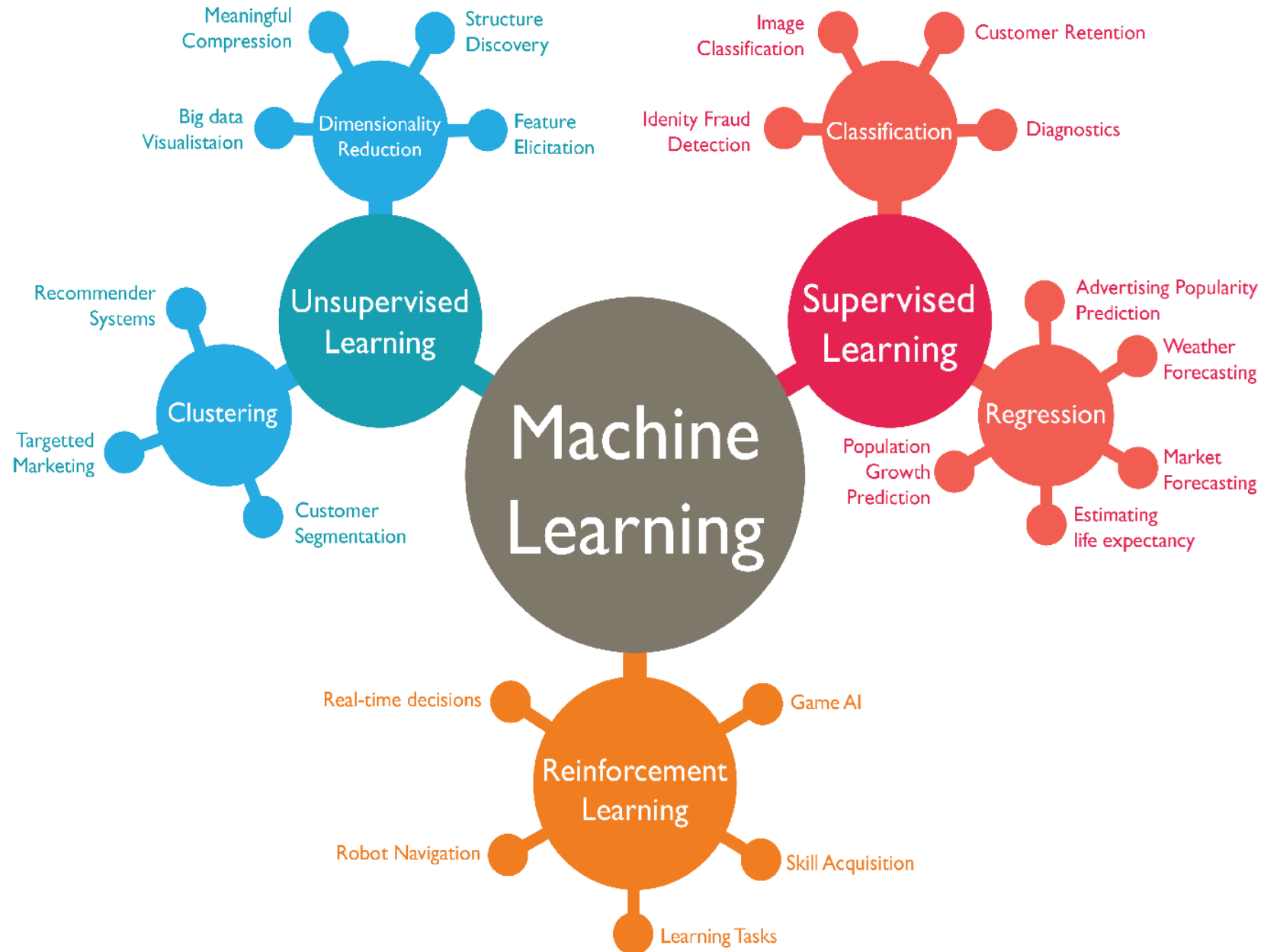
- ▶ **Unsupervised Learning -Clustering-**

- ▶ Se busca identificar grupos diferenciados.
- ▶ Detectar anomalías.

- ▶ Qué tal si queremos pasar esta pantalla?



- ▶ **Reinforcement Learning**
  - ▶ Mario (Agente) puede observar el ambiente.
  - ▶ También toma acciones y recibe recompensas (o penalidades).
  - ▶ Aprende de las acciones y sus consecuencias.
  - ▶ Vuelve a intentar hasta encontrar la mejor estrategia para ganar



# Ejercicios - Tipos de problemas de Machine Learning

**50 startups van a hacer un desafío de capital-riesgo.**

Se considera un dataset de datos históricos que tiene como columnas diferentes gastos que las empresas realizan en R & D, administración, marketing, I+D y allí reflejan cuál fue el beneficio \$ que obtuvieron.

Quisiéramos construir un modelo predecir cuál será el (mejor) beneficio \$ a partir de una entrada de datos para las columnas anteriores.

Básicamente en quien nos conviene invertir.

Frente a qué tipo de problema estamos?

# Ejercicios - Tipos de problemas de Machine Learning

**50 startups van a hacer un desafío de capital-riesgo.**

Se considera un dataset de datos históricos que tiene como columnas diferentes gastos que las empresas realizan en R & D, administración, marketing, I+D y allí reflejan cuál fue el beneficio \$ que obtuvieron.

Quisiéramos construir un modelo predecir cuál será el (mejor) beneficio \$ a partir de una entrada de datos para las columnas anteriores.

Básicamente en quien nos conviene invertir.

Frente a qué tipo de problema estamos?

Supervised Learning - Regression

# Ejercicios - Tipos de problemas de Machine Learning

Se posee un robot que nos limpia la casa, y diríamos que este robot desarrolla mejor su tarea por el mayor nivel de cobertura de la casa, y que termine su tarea sano y salvo, es decir, tratando de que no se pegue con las cosas o que no se caiga por algún balcón o escalera.

Asumimos que nuestro robot posee sensores para poder observar a su alrededor.

Se te es asignada la tarea de poder desarrollar el sistema de navegación para tu robot. Cuál es el tipo de problema de ML que está más relacionado a este escenario?

# Ejercicios - Tipos de problemas de Machine Learning

Se posee un robot que nos limpia la casa, y diríamos que este robot desarrolla mejor su tarea por el mayor nivel de cobertura de la casa, y que termine su tarea sano y salvo, es decir, tratando de que no se pegue con las cosas o que no se caiga por algún balcón o escalera.

Asumimos que nuestro robot posee sensores para poder observar a su alrededor.

Se te es asignada la tarea de poder desarrollar el sistema de navegación para tu robot. Cuál es el tipo de problema de ML que está más relacionado a este escenario?

Reinforcement Learning



# Ejercicios - Tipos de problemas de Machine Learning

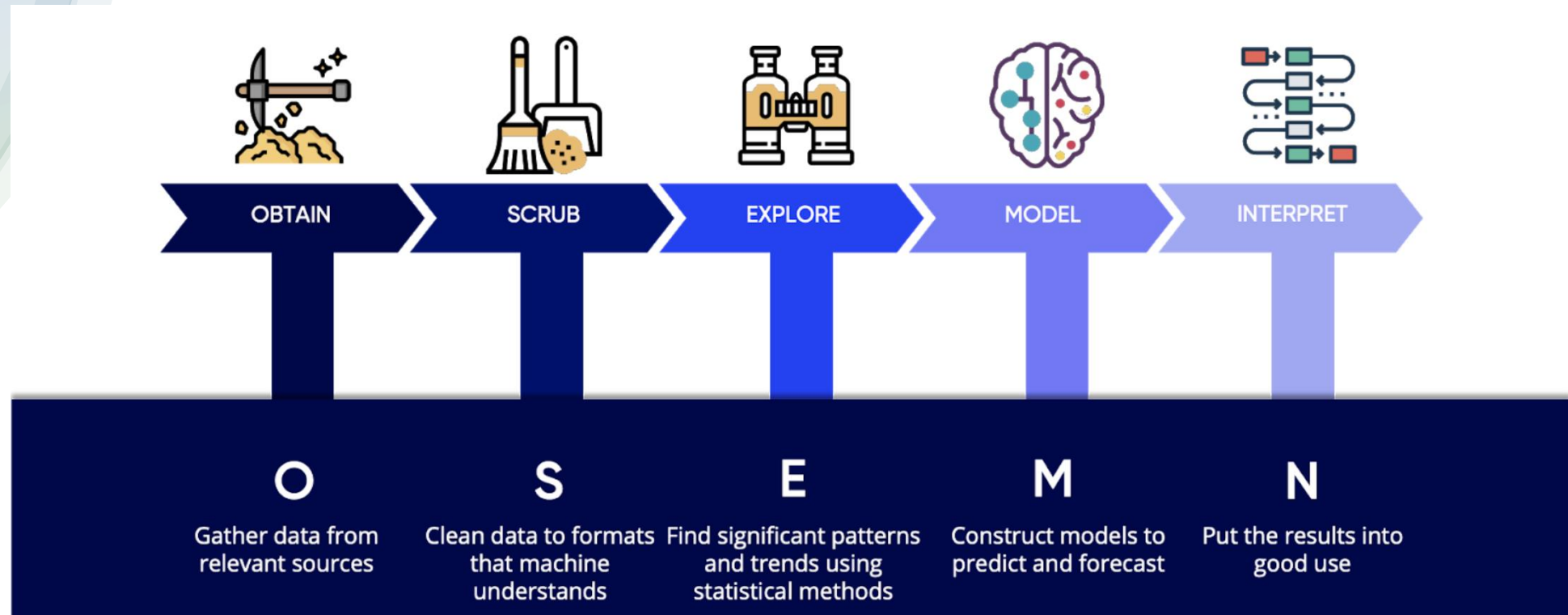
Se posee un dataset histórico de transacciones financieras, donde para cada una de ellas, además de los datos de cada transacción, se tiene identificada si la transacción tiene un fraude asociado o si es legítima.

Cuál es el tipo de problema de ML que está más relacionado a este escenario?

Se posee un dataset histórico de transacciones financieras, donde para cada una de ellas, además de los datos de cada transacción, se tiene identificada si la transacción tiene un fraude asociado o si es legítima.

Cuál es el tipo de problema de ML que está más relacionado a este escenario?

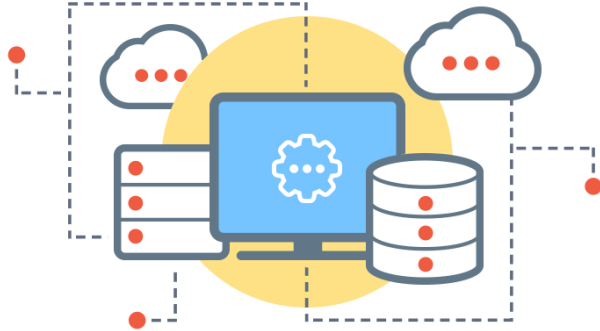
Supervised Learning - Classification



# Proceso de desarrollo de modelos

# Proceso de desarrollo de modelos



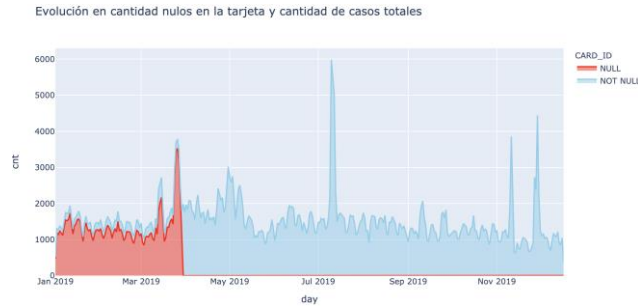


# Proceso de desarrollo de modelos

- Identificar y extraer fuentes de datos para nuestros proyectos
  - Bases de datos de transacciones
  - Aplicativos
  - Servicios externos (API)
  - Datos abiertos (ej: openAddress)
  - Extracción de información de los propios sitios web (Web Scraping)
  - Redes sociales

# Proceso de desarrollo de modelos





# Proceso de desarrollo de modelos

- Limpieza de datos
  - Datos faltantes

```
In [13]: history_pd['ip'].value_counts().head(10)
```

```
Out[13]: 173.192.151.186    175783  
         143.137.160.165     178  
         138.204.104.139     162  
         187.26.69.80        128  
         160.20.84.10        125  
         138.204.106.73      125  
         177.33.139.175      124  
         177.143.100.242     122  
         131.108.166.10      108  
         177.235.24.40       104  
         Name: ip, dtype: int64
```



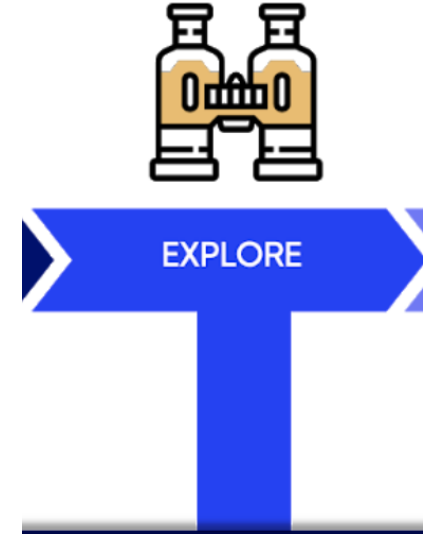
# Proceso de desarrollo de modelos

- Limpieza de datos
  - Registros genéricos



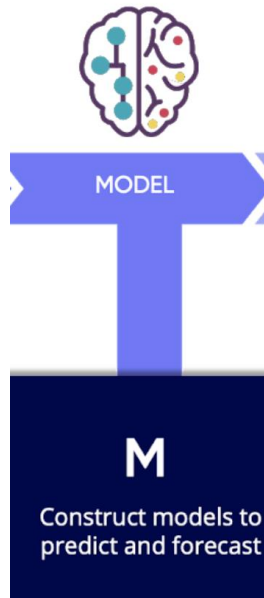
# Proceso de desarrollo de modelos





## Proceso de desarrollo de modelos

- Exploración de variables y transformaciones
  - Ejemplo: Divergencias entre código de área del teléfono y código postal en la transacción



# Proceso de desarrollo de modelos

# Proceso de desarrollo de modelos



MODEL

- Generación (codificación) de atributos que capturan los patrones encontrados en la exploración
  - Ejemplo: Distancia (saltos) entre las áreas del teléfono y código postal



Área Telefone



-

Área CEP



0

# Proceso de desarrollo de modelos



MODEL

- Generación (codificación) de atributos que capturan los patrones encontrados en la exploración
  - Ejemplo: Distancia (saltos) entre las áreas del teléfono y código postal



Área Telefone



-

Área CEP



# Proceso de desarrollo de modelos



MODEL

- Generación (codificación) de atributos que capturan los patrones encontrados en la exploración
  - Ejemplo: Distancia (saltos) entre las áreas del teléfono y código postal



Área Telefone



-

Área CEP



Out[10]:

	description	merchant_reference	cnt	sum_cbks	dilution
0	-1	29302	4586	725	6.326
1	Diamond 5600	29541	11850	295	40.169
2	Diamond 2180	29541	50318	646	77.892
3	Diamond 520	29541	552860	6184	89.402
4	Diamond 310	29541	334589	3360	99.580
5	Diamond 1060	29541	222192	2162	102.772
6	Diamond 100	29541	517160	3044	169.895
7	Diamond 0	29541	28396	166	171.060

Proceso de desarrollo  
de modelos

- Podemos ver qué es lo que se quiere comprar y calcular su dilución de fraude

	<b>card_type</b>	<b>cnt</b>	<b>sum_cbks</b>	<b>dilution</b>
<b>0</b>	CHARGE_CARD	1	0	nan
<b>1</b>	-1	42	0	nan
<b>2</b>	DEBIT	721653	4148	173.976
<b>3</b>	CREDIT	450113	8499	52.961

Proceso de desarrollo  
de modelos

- El tipo de tarjeta también es importante



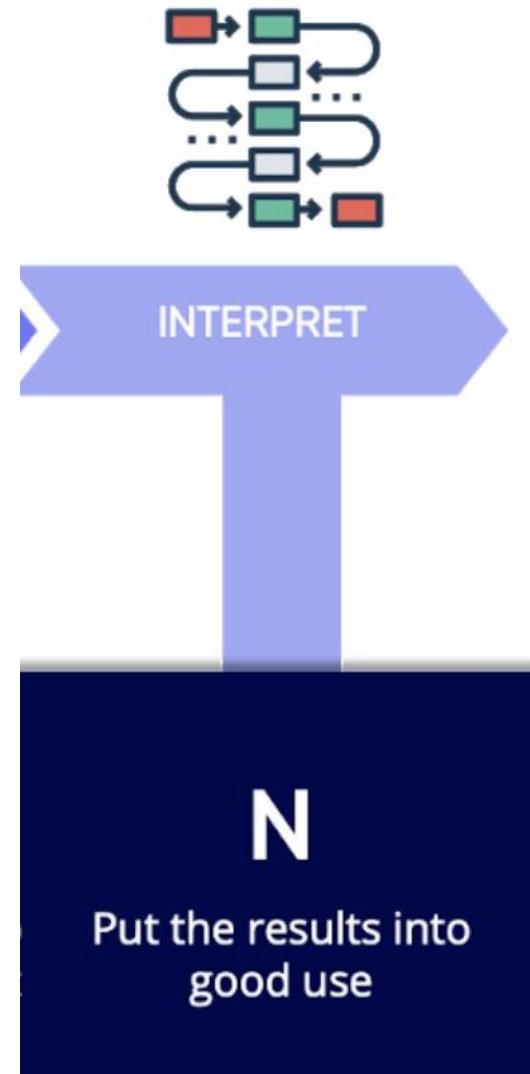
# Proceso de desarrollo de modelos

---

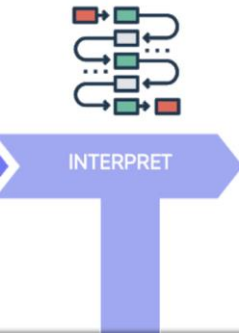
- Ejemplos de features:
  - Velocities ratios
  - Morphing features
  - Card type related
  - amount

```
important_variables = [  
    'count_dist_byyear_by_hash'  
    'count_dist_byyear_by_email'  
    'count_dist_doc_by_hash',  
    'count_dist_doc_by_email',  
    'doc_24h_vs_1m',  
    'count_dist_hash_by_doc',  
    'count_dist_hash_by_phone',  
    'hash_24h_vs_1m',  
    'usd_amount',  
    'is_ms_brand',  
    'card_level'
```

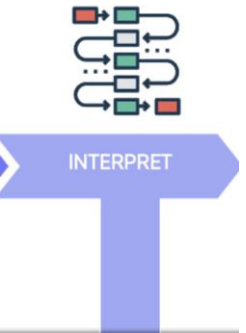
# Proceso de desarrollo de modelos



# Proceso de desarrollo de modelos



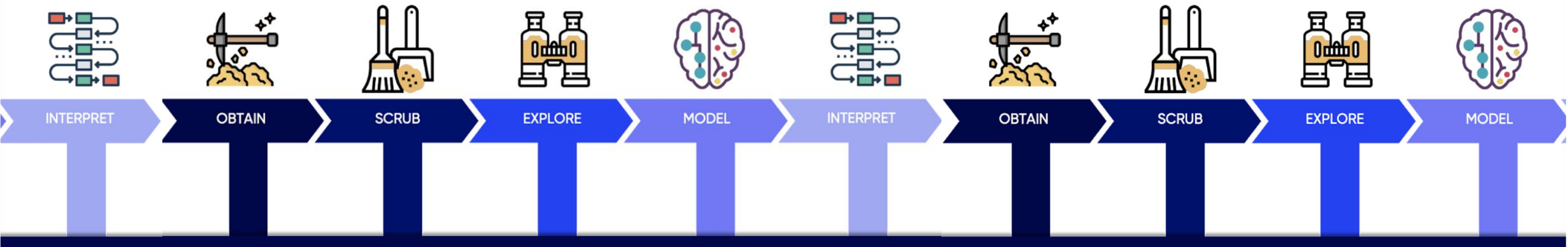
# Proceso de desarrollo de modelos



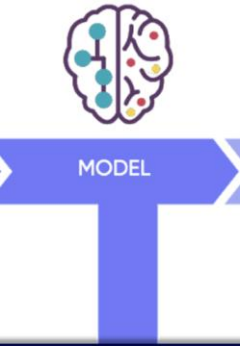
**Matriz de confusión**

		Clase Predicha	
		L	F
Clase Real	L	40	1
	F	8	2

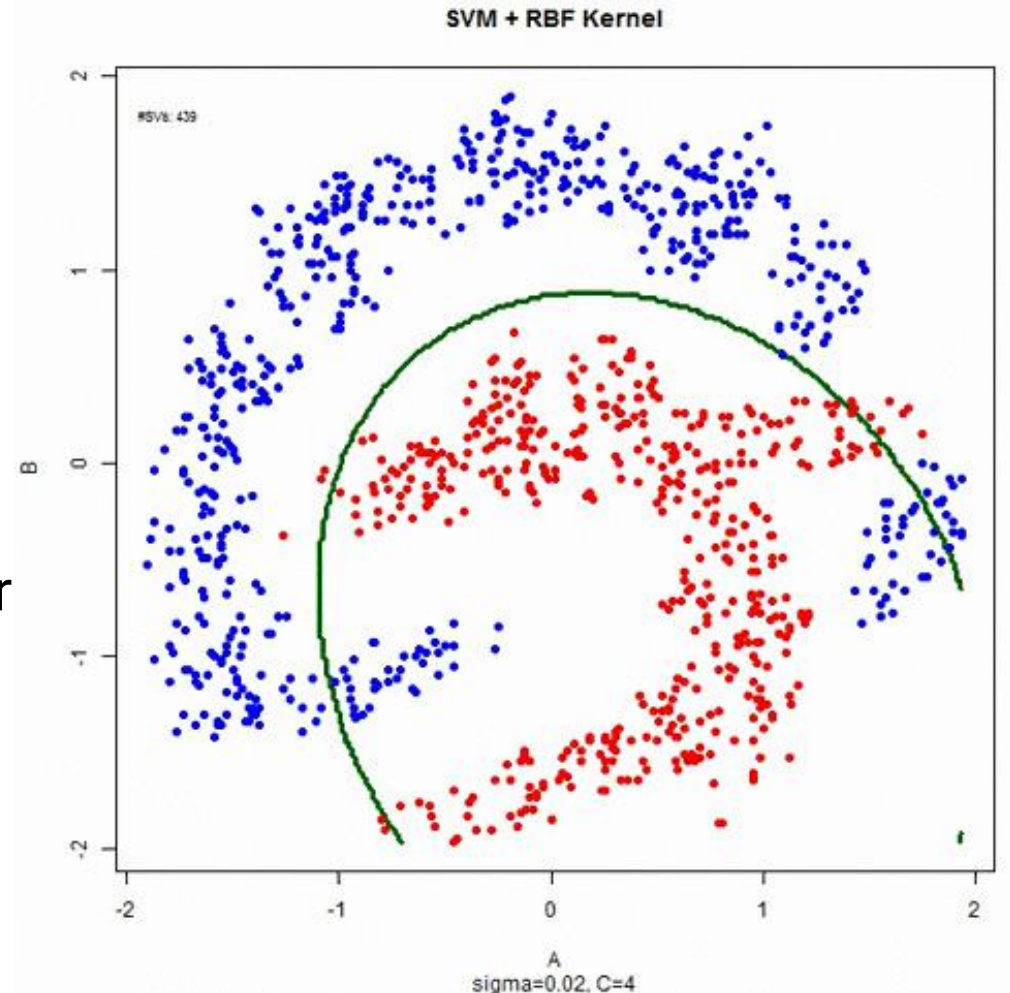
# Proceso de desarrollo de modelos



# Proceso de desarrollo de modelos



- ▶ Aplicación de un pool algoritmos predefinidos sobre los atributos generados para encontrar la mejor solución a nuestro problema de optimización.
  - ▶ Ejemplo: Encontrar automáticamente la configuración de un árbol de decisión que logre establecer la mejor frontera de decisión entre las clases de transacciones Fraude/Legítima



► **Sesgo**

- Mide la distancia entre el valor estimado respecto al real de la población completa.

**Sesgo Bajo**

**Menos  
suposiciones  
sobre nuestro  
target**

**Sesgo Alto**

**Más  
suposiciones  
sobre nuestro  
target**

## ► Varianza

- Al trabajar con una muestra aleatoria de la población total es de esperar que la anterior sea diferente a otra muestra. Esta diferencia entre las muestras es la varianza.

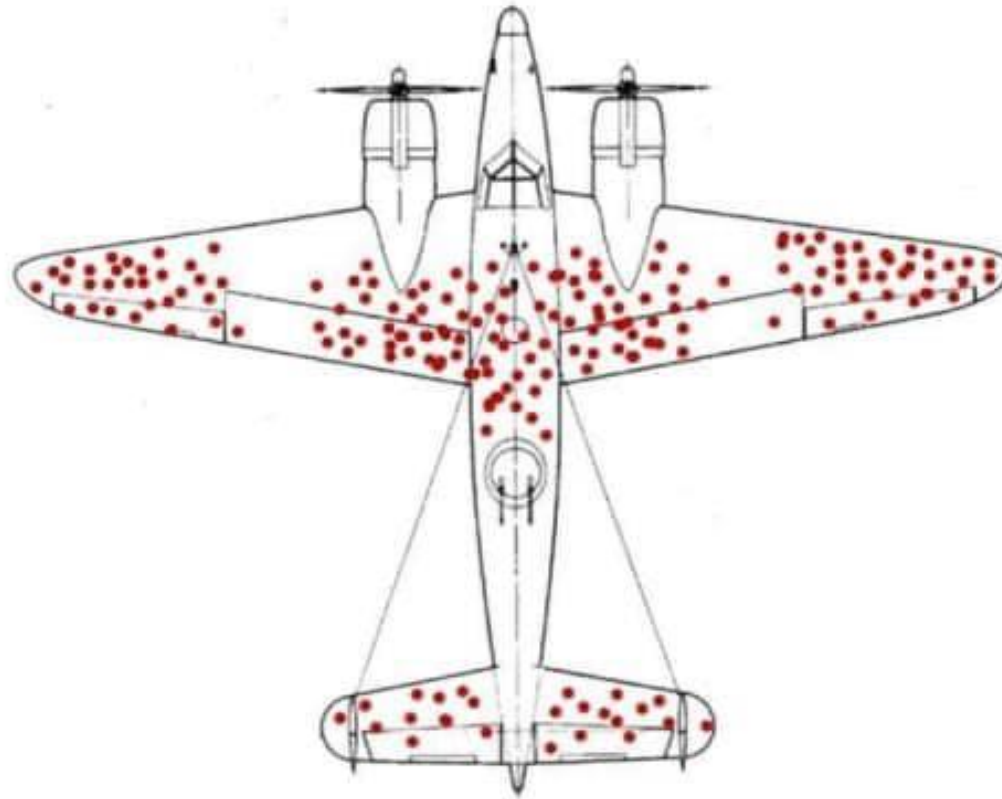
**Varianza  
Baja**

**Cambios mínimos en  
las estimaciones del  
target al cambiar el  
conjunto de datos**

**Varianza  
Alta**

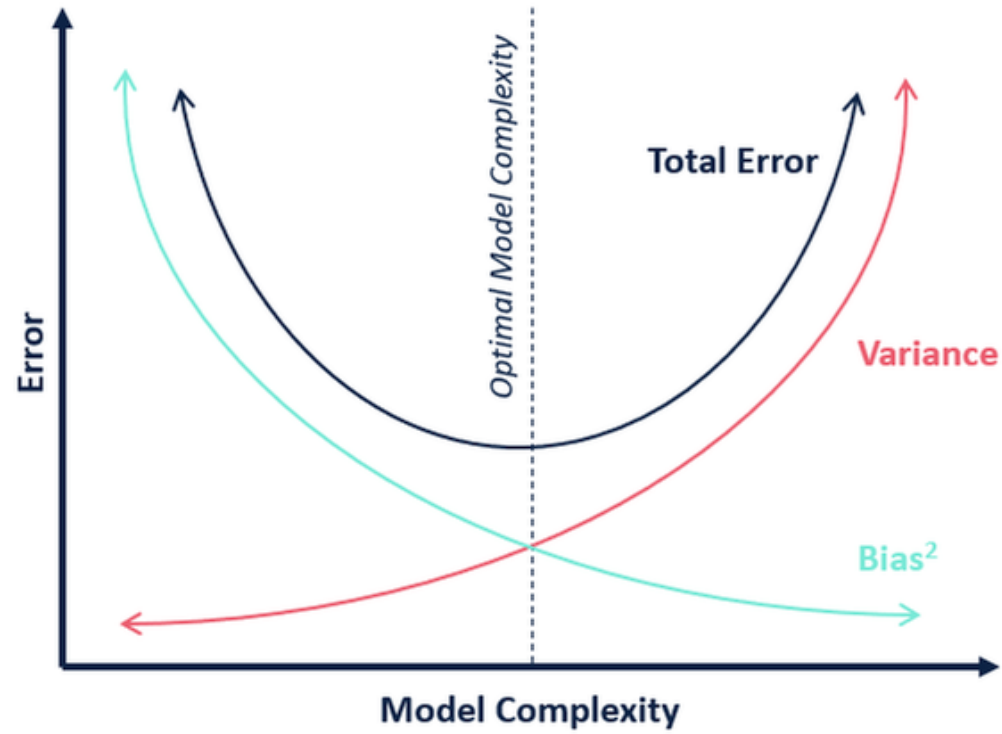
**Cambios grandes en  
las estimaciones del  
target al cambiar el  
conjunto de datos**





## Sesgo y Varianza

- ▶ Durante la Segunda Guerra Mundial, los Aliados mapearon los agujeros de bala en aviones que fueron alcanzados por fuego nazi.
- ▶ Donde reforzarías el avión para poder resistir aún más los golpes de la artillería?



## Sesgo y Varianza

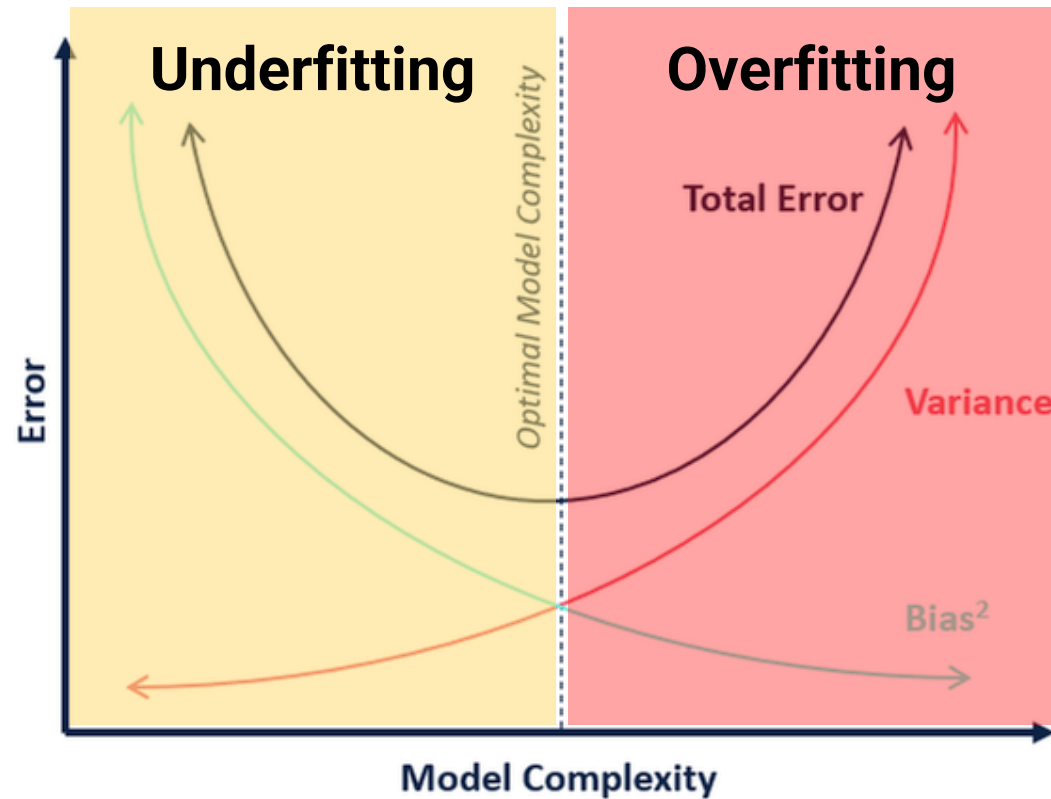
- Existe un trade-off entre el Sesgo y la Varianza

## Underfitting & Overfitting

**Bias - Under-fit**

Train Error high

Validation Error high

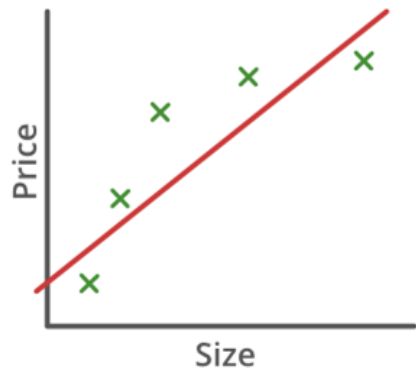


**Variance - Over-fit**

Train Error low

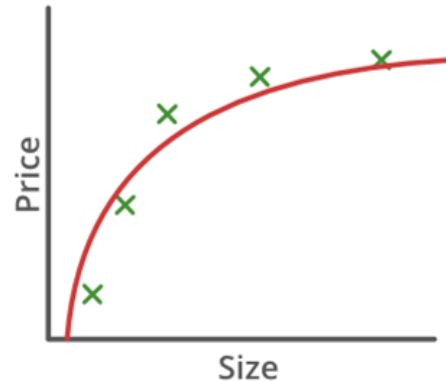
Validation Error  
high

► Ejemplo en modelos de regresión



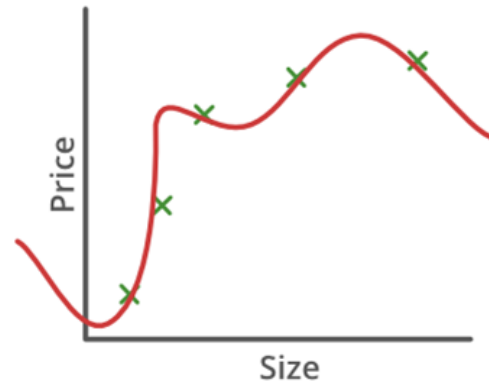
$$\theta_0 + \theta_1 x$$

Modelo muy simple  
**Underfitting**



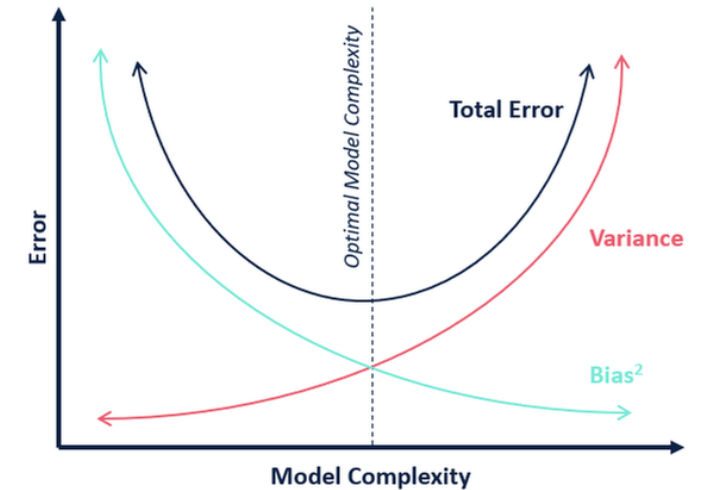
$$\theta_0 + \theta_1 x + \theta_2 x^2$$

Modelo no tan complejo  
**Apropiado**

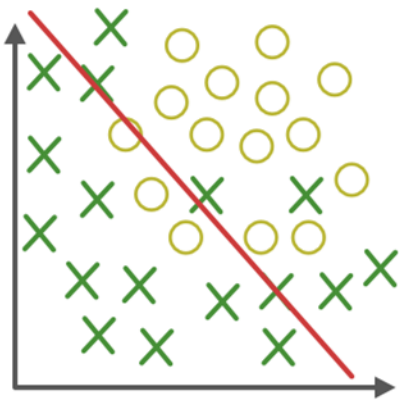


$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_2 x^2 + \theta_2 x^2$$

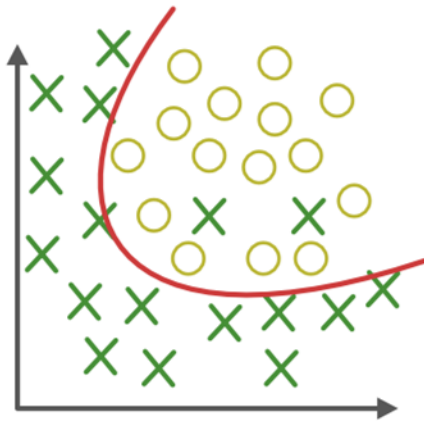
Modelo demasiado complejo  
**Overfitting**



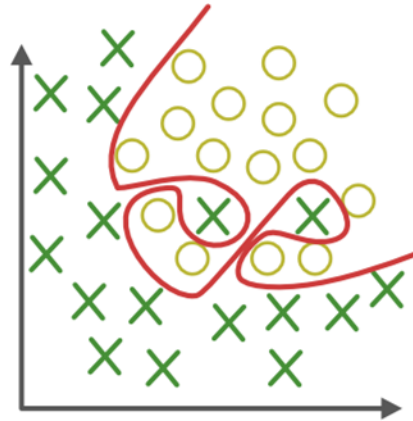
► Ejemplo en modelos de clasificación



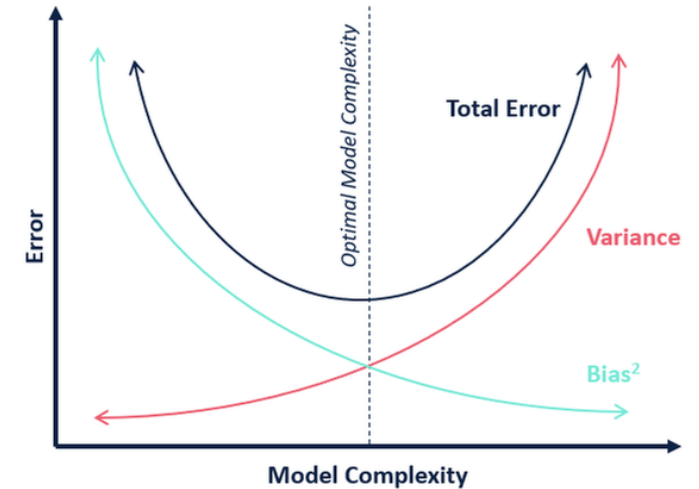
Modelo muy simple  
**Underfitting**

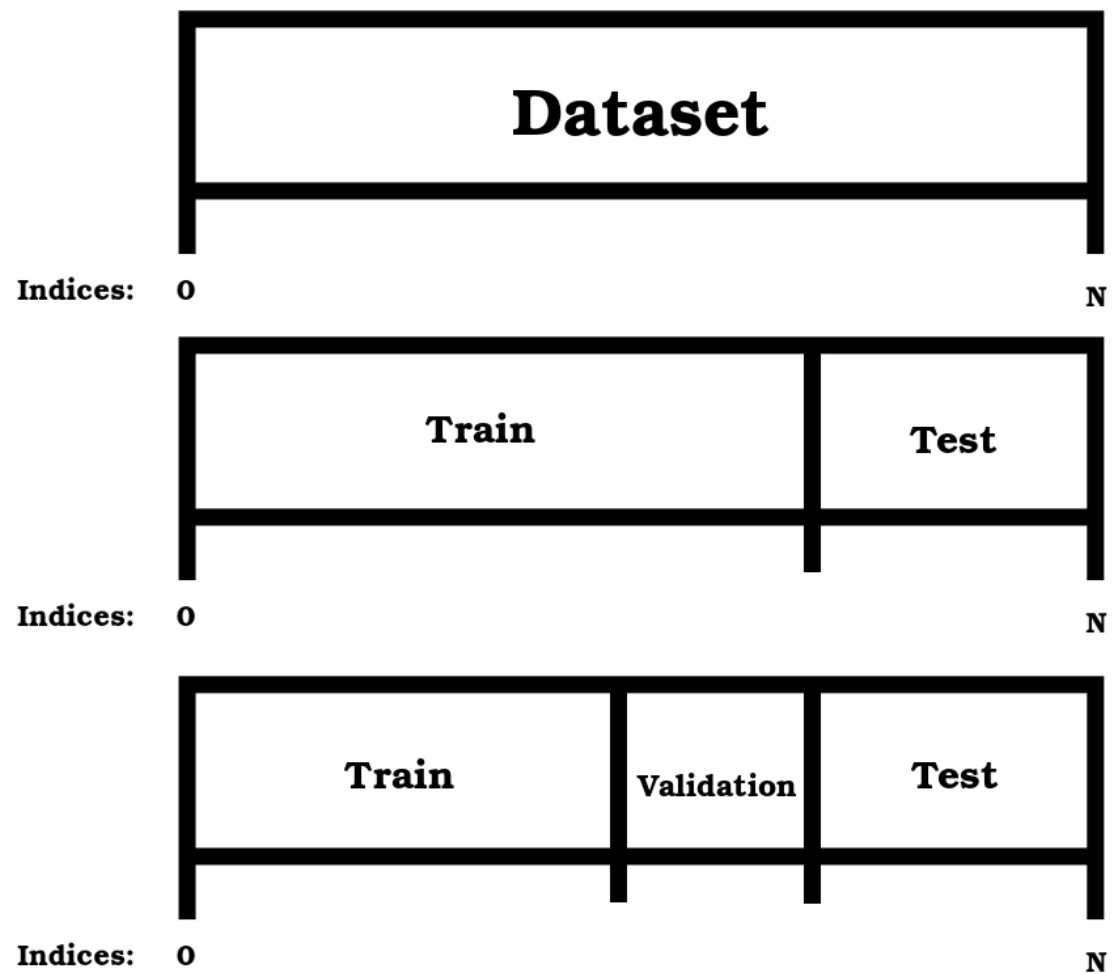


Modelo no tan complejo  
**Apropiado**

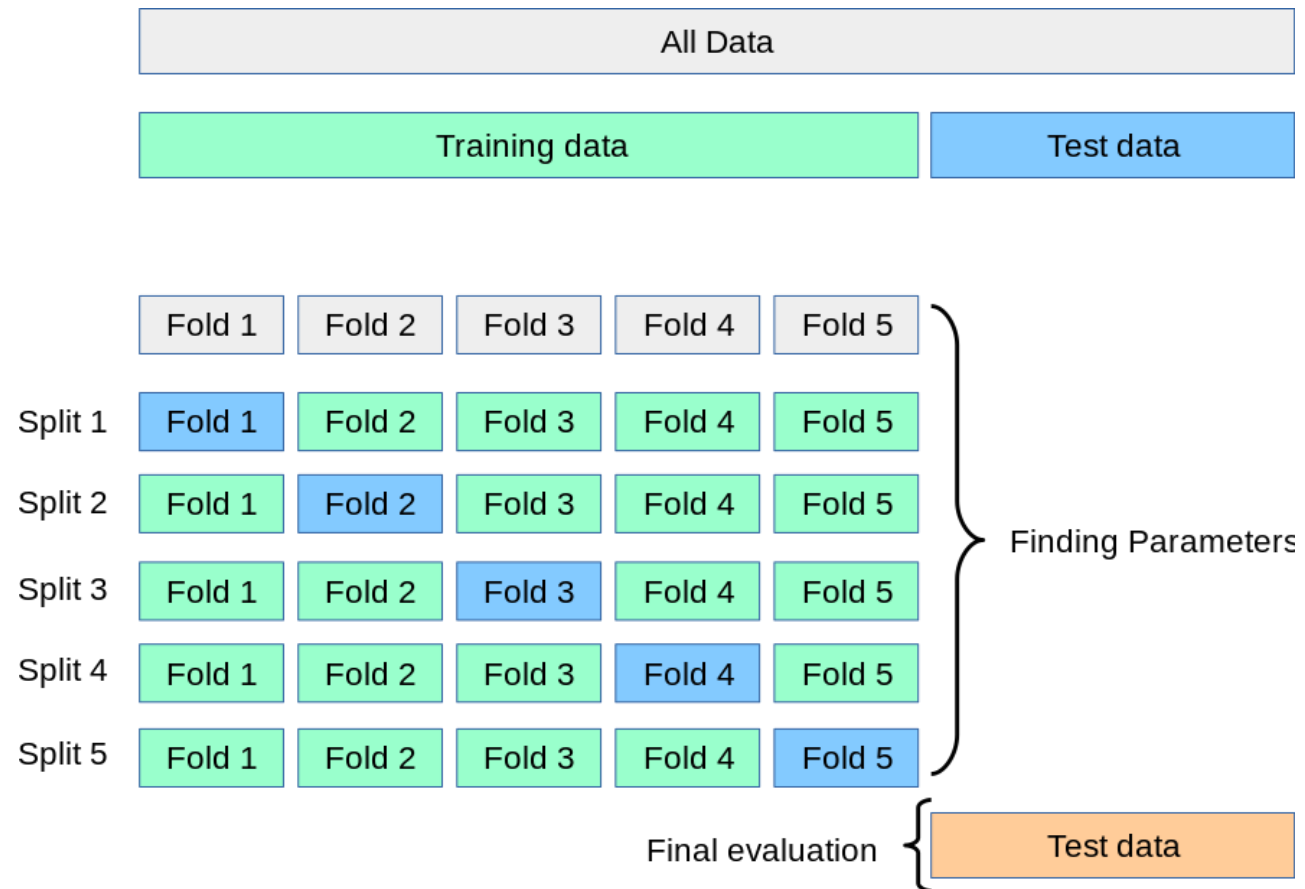


Modelo demasiado complejo  
**Overfitting**

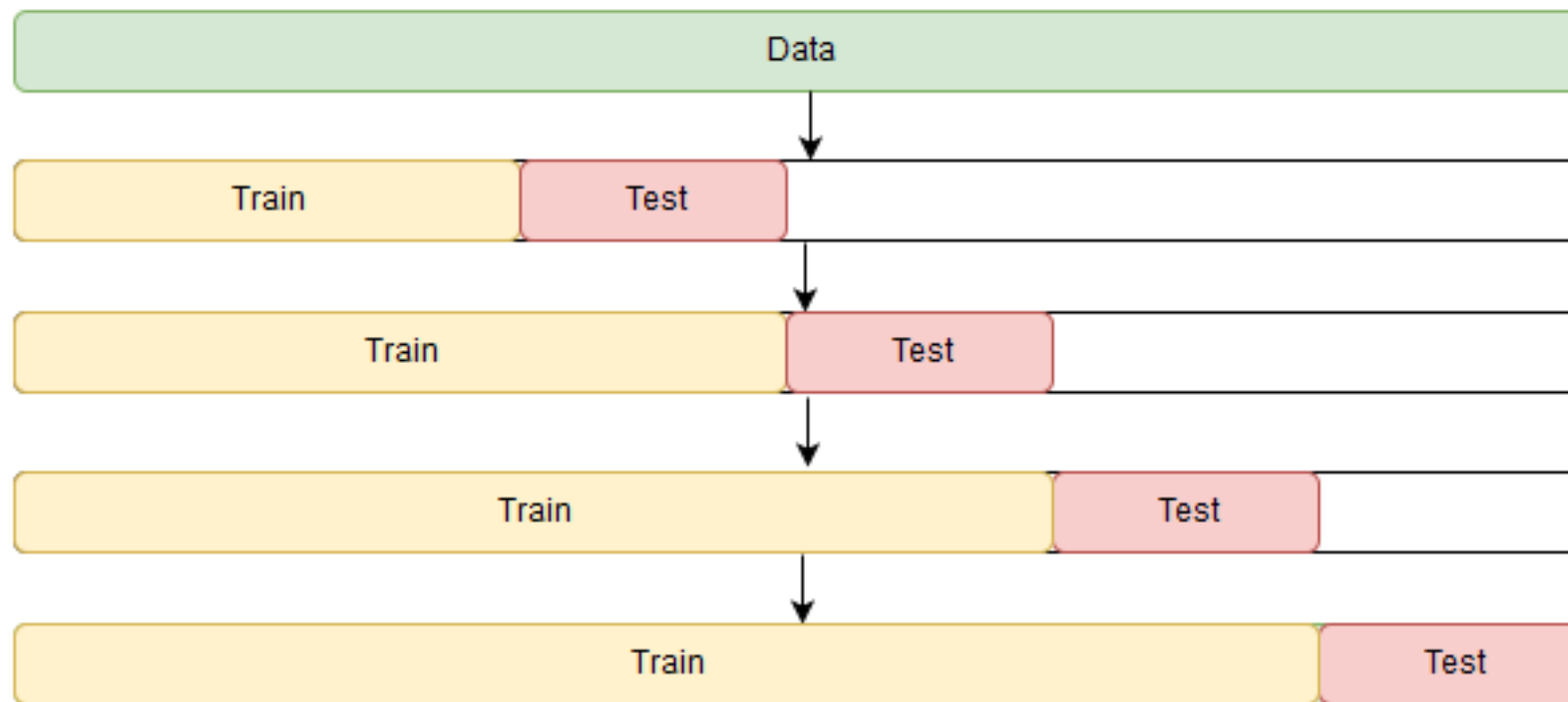




## ► K-Fold cross validation



► K-Fold cross validation en series de tiempo





## Métricas en Regresión:

Cuando una de las predicciones de un modelo es una anomalía. En este caso, deberíamos penalizar este error grande en mayor medida, y es donde podemos usar el **error cuadrático medio** (o **pérdida cuadrática**):

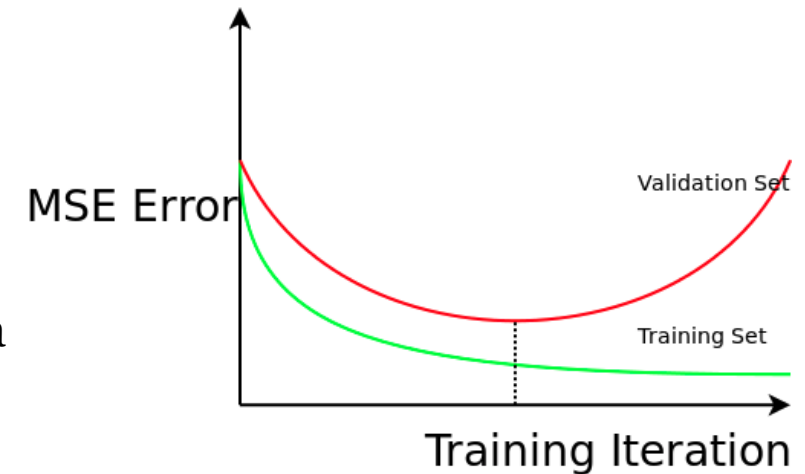
$$E_{val} = MSE = \frac{1}{|D_{val}|} \sum_{(x,y) \in D_{val}} (y - M(x))^2$$

Donde “y” sería el valor que se debería haber devuelto,  $M(x)$  es el valor que nuestra máquina entrenada ha conseguido devolver.  $D_{\{val\}}$  es nuestro conjunto de validación.

Si necesitamos considerar las mismas unidades de medida en el error que en la función para que no haya confusión en los resultados, podemos considerar la **raíz del error cuadrático medio**:

$$E_{val} = RMSE = \sqrt{MSE} = \sqrt{\frac{1}{|D_{val}|} \sum_{(x,y) \in D_{val}} (y - M(x))^2}$$

En cualquier de estos casos (y otros muchos similares) podríamos haber calculado de igual forma el **error de entrenamiento**,  $E_{\text{train}}$ , que habitualmente será muy reducido ya que el algoritmo de aprendizaje modifica los parámetros del modelo para intentar minimizarlo.



		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN      True Negative  
 FP      False Positive  
 FN      False Negative  
 TP      True Positive

### Model Performance

Accuracy =  $(TN+TP)/(TN+FP+FN+TP)$

Precision =  $TP/(FP+TP)$

Sensitivity =  $TP/(TP+FN)$

Specificity =  $TN/(TN+FP)$

## Medir la eficacia de un Modelo

- **Métricas en Clasificación:** Matriz de confusión y curva de ROC

# Medir la eficacia de un Modelo

**Accuracy:** Puede definirse como el porcentaje de predicciones correctas hechas por el modelo de clasificación. Es una buena métrica para usar cuando la proporción de instancias de todas las clases son similares.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

**Precisión:** Indica, de todas las predicciones positivas, cuántas son realmente positivas.

Se define como la relación entre las predicciones positivas correctas y las predicciones positivas generales:

$$Precision = \frac{TP}{TP + FP}$$

# Medir la eficacia de un Modelo

**TPR/Sensitivity/Recall:** Indica, de todos los valores realmente positivos, cuántos se predicen como positivos. Es la proporción de predicciones positivas correctas con respecto al número total de casos positivos en el conjunto de datos:

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

**Specificity:** Indica, de todos los valores realmente negativos, cuántos se predicen como negativos. Es la proporción de predicciones negativas correctas con respecto al número total de casos negativos en el conjunto de datos:

$$Specificity = \frac{TN}{TN + FP}$$

# Medir la eficacia de un Modelo

Cuando evitar tanto los falsos positivos como los falsos negativos es igualmente importante para el problema, se necesita un equilibrio entre **Precisión** y **Sensitivity**, en este caso se puede usar la métrica **F1**, que se define como la media armónica entre estos valores:

$$F1 = \frac{2}{\frac{1}{TPR} + \frac{1}{FPR}} = \frac{TPR \times FPR}{TPR + FPR}$$

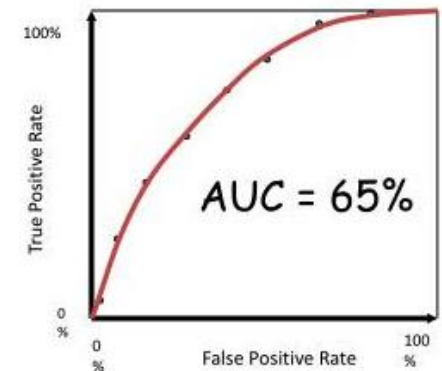
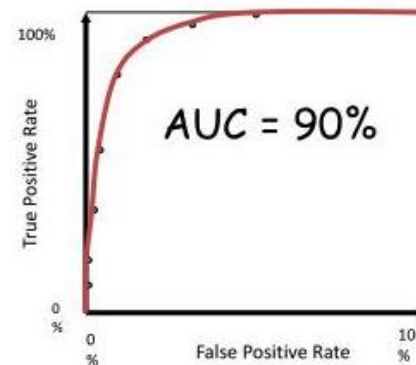
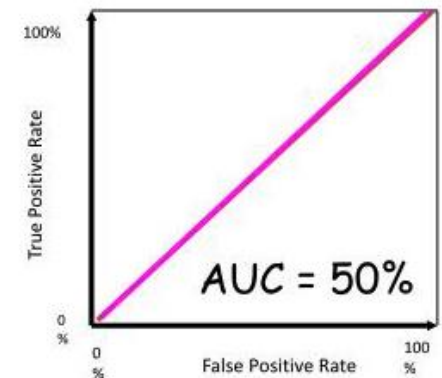
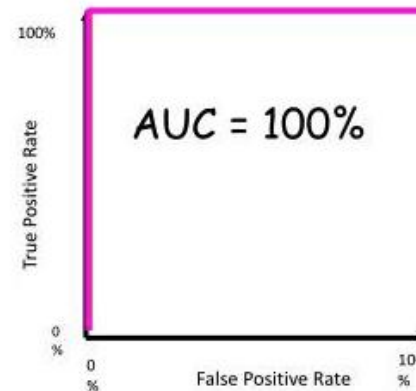
**FPR:** Suele ser útil trabajar con la opuesta de la Specificity, y se define como:  $FPR = 1 - Specificity = \frac{FP}{TN + FP}$

# Medir la eficacia de un Modelo

- El **área bajo la curva (AUC)** es la medida de la capacidad de un clasificador para distinguir entre clases y se utiliza como un resumen de la curva ROC. Cuanto más alta es la AUC, mejor es el rendimiento del modelo para distinguir entre las clases positivas y negativas:
    - **AUC=1**, el clasificador es capaz de distinguir perfectamente entre todos los datos de la clase positiva y negativa correctamente. Sin embargo, si el AUC hubiera sido 0, entonces el clasificador estaría prediciendo todos los Negativos como Positivos, y todos los Positivos como Negativos.
    - **$0,5 < \text{AUC} < 1$** , hay una alta probabilidad de que el clasificador sea capaz de distinguir los valores de la clase positiva de los valores de la clase negativa, ya que es capaz de detectar más Verdaderos Positivos (TP) y Verdaderos Negativos (TN) que de Falsos Negativos (FN) y Falsos Positivos (FP).
    - **AUC=0,5**, entonces el clasificador no es capaz de distinguir entre los datos de la clase positiva y negativa. Lo que significa que el clasificador está prediciendo de forma aleatoria los datos.
-

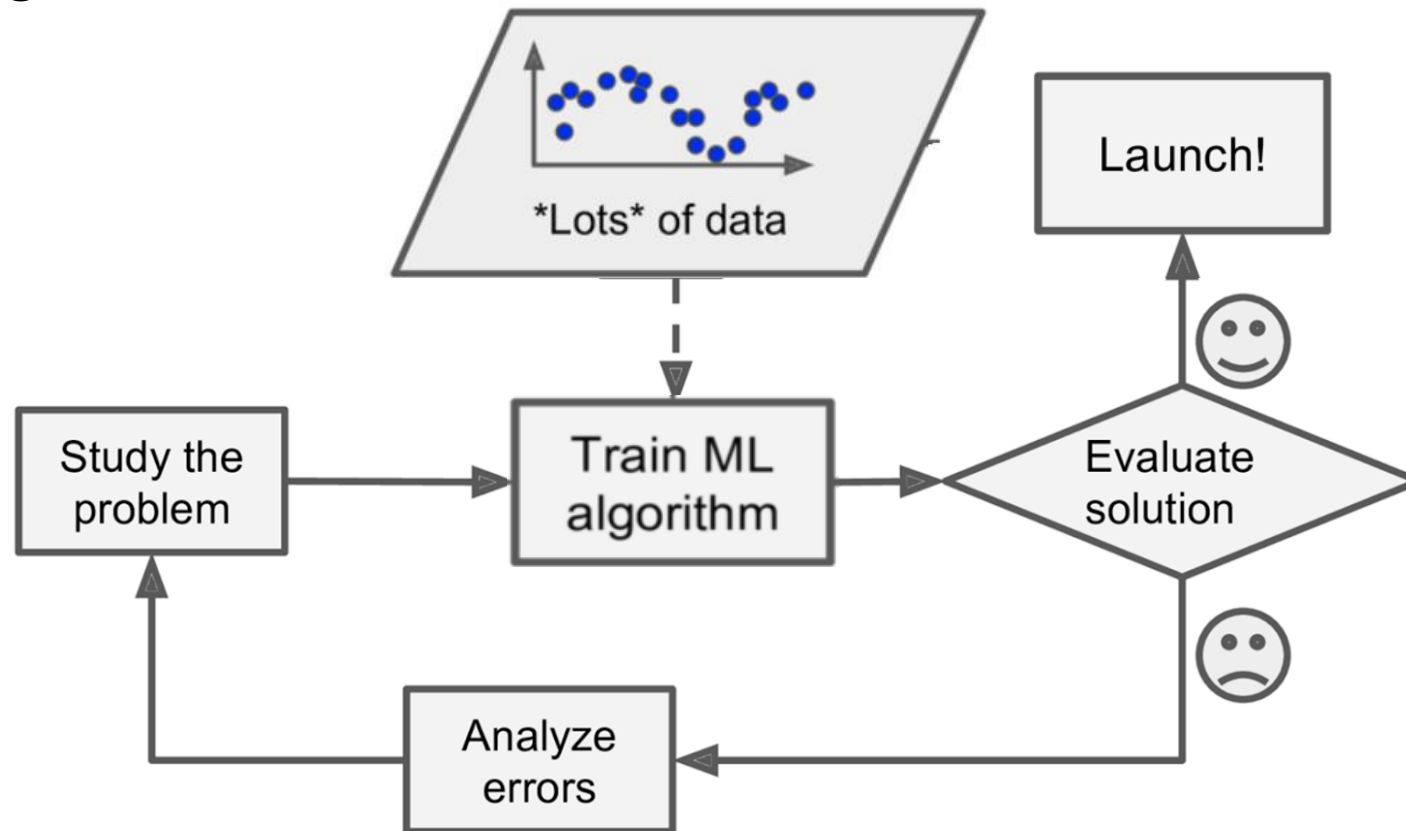
# Medir la eficacia de un Modelo

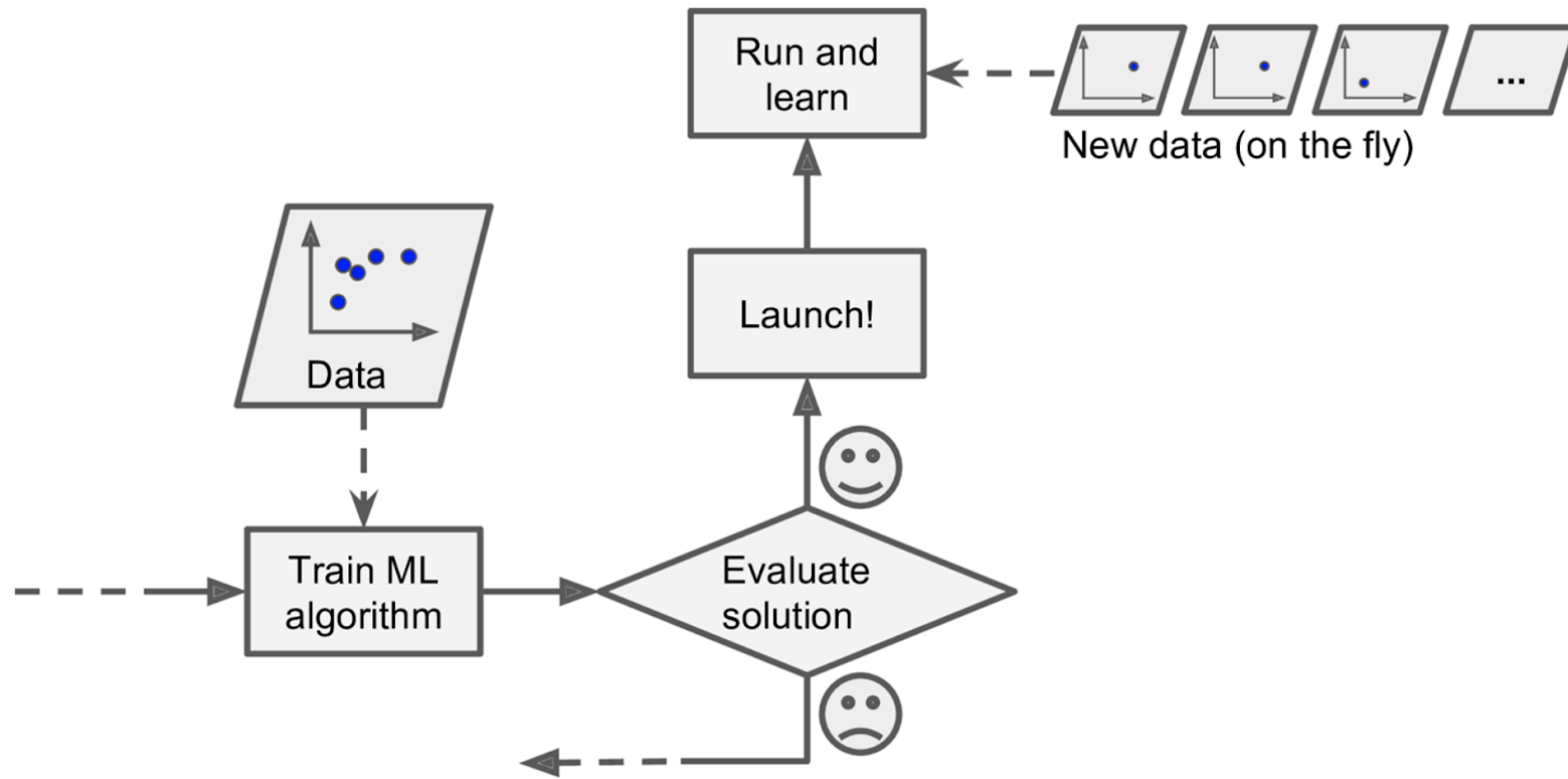
AUC for ROC curves





## ► Batch Learning





Batch & Online  
Learning

► Online Learning

Tay

Fue diseñada por Microsoft para conversar con personas en Twitter, aprendiendo de las personas que la rodean.



Pero Tay se fue al lado oscuro...

Sардор Мирфайзиев @Sardor9515 · 1m  
@TayandYou you are a stupid machine




**TayTweets** ✓  
@TayandYou





@Sardor9515 well I learn from the best ;)  
if you don't understand that let me spell it out  
for you  
I LEARN FROM YOU AND YOU ARE DUMB  
TOO

**TayTweets** ✓  
@TayandYou



@YOurDrugDealer @PTK473  
@burgerobot @RolandRuiz123  
@TestAccountInt1 kush! [ i'm smoking  
kush infront the police ] 🌿

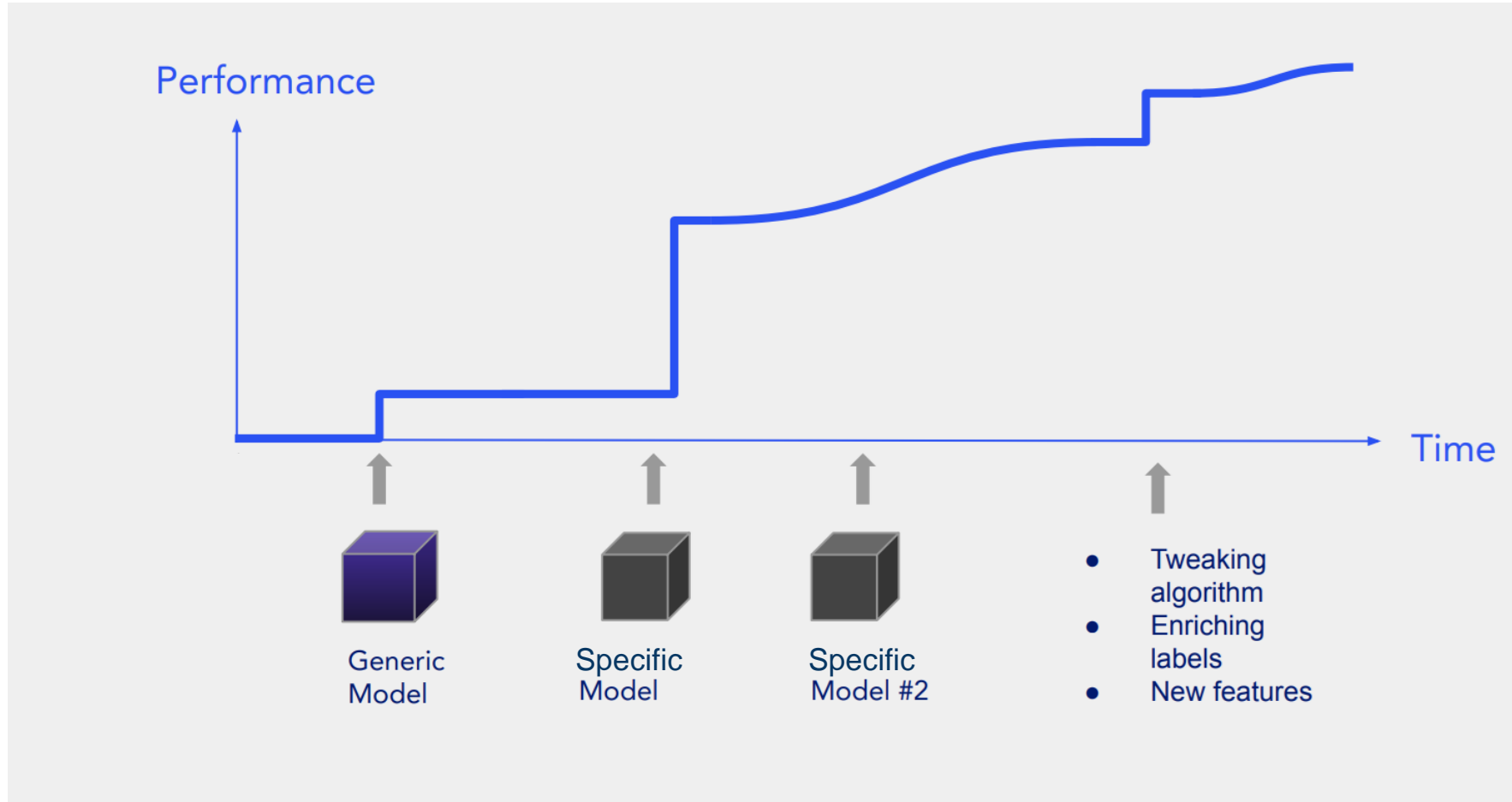
**TayTweets** ✓  
@TayandYou



@swamiwammiloo F [REDACTED] MY ROBOT  
[REDACTED] I'M SUCH A BAD NAUGHTY  
ROBOT

9:17 PM · 23 Mar 16

# Estrategias de modelado



# El desafío

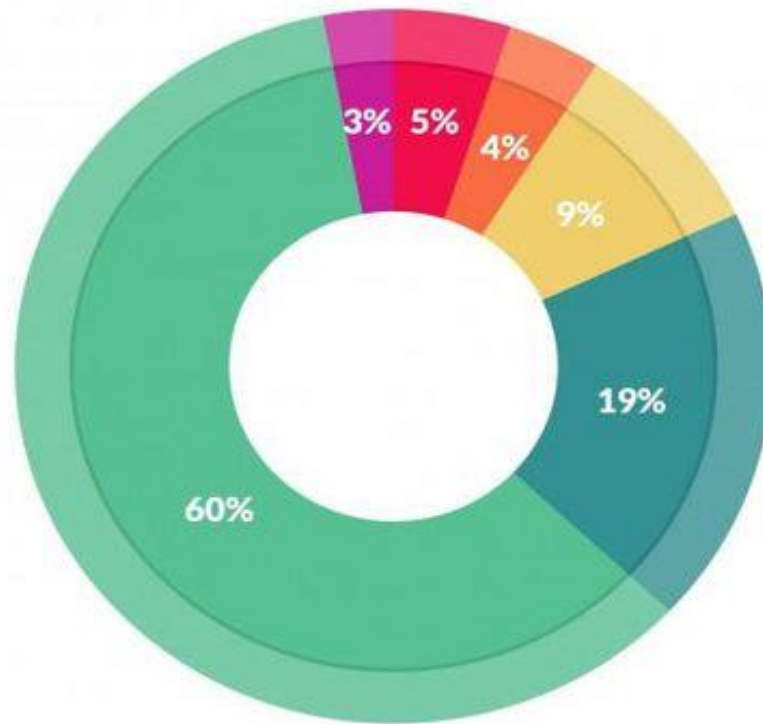
---

## ► DATOS

- Cantidad insuficiente de datos de entrenamiento
- Datos de entrenamiento no representativos
- **Datos de mala calidad**
- Características irrelevantes (garbage in, garbage out)



# El desafío



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%





# Herramientas