# INTRODUCCIÓN MACHINE LEARNING

Tomás Fontecilla

12 de agosto de 2022

# PARA TENER EN CUENTA

## REGLAS BÁSICAS

- Se discutirá los algoritmos propios de Machine Learning desde visión usuaria.
- El software preferente en este curso es Python.
- Me pueden contactar por mail: tfontecillac@udd.cl
- Toda pregunta es bienvenida.
- toda duda y comentario sirve.
- Por modalidad, usaré pausas para revisar comentarios online

# Tomás Fontecilla

Matemático Estadístico de Pontificia Universidad Católica de Chile, Magíster en Estadística, con 15 años de experiencia en data science. Seré su profesor para este curso.

# PROGRAMA DEL CURSO

**Fundamentos**

- ¿ Qué es Machine Learning?
- Introducción a scikit-learn.
- Hiperparámetros y validación de modelos.
- Ingeniería de Características (Feature Engineering).

**Machine Learning - Aprendizaje Supervisado**

a) Naive Bayes
b) Análisis Discriminante
c) Regresión Logística
d) *Support Vector Machine*
e) Evaluación de modelos de clasificación
f) Estrategias para datos desbalanceados
g) *k-nearest neighbors*
h) Modelos de árboles de decisión y *random forest*
i) Boosting: AdaBoost y XGBoost

**Machine Learning - Aprendizaje No Supervisado**

a) Análisis de componentes principales
b) *K-means clustering*
c) Clustering jerárquico
d) *Gaussian Mixtures* (mezclas gaussianas)
e) *Kernel Density Estimation*
f) Escalado y variables categóricas

# ¿ Qué es Machine Learning?

¿ Qué es Machine Learning?
¿ Cómo se define la disciplina?
¿ Cuáles son sus principales aplicaciones?

# Principales conceptos asociados

- Estadística
- Matemática
- Minería de datos
- Inteligencia artificial
- Inteligencia de negocios
- Predicción

# Business Intelligence, Inteligencia de Negocios

Extracting useful information from large data sets. (Hand et al., 2001)
*Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules.* (Berry and Linoff, 1997, p. 5)

# La definición que más se asemeja hoy es:

[Data Mining is]*the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories.Data mining employs pattern recognition technologies,as well as statistical and mathematical techniques.* (Garnet Group, 2010)

# Pero, ¿de qué trata la minería de datos? ¿Existe alguna metdología estándar de trabajo?

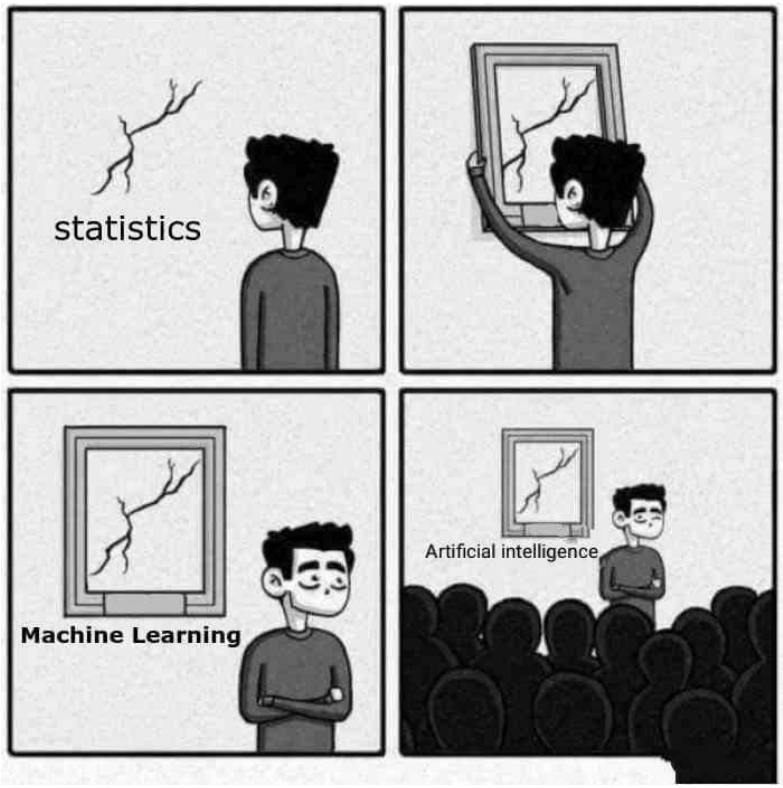

¿ Cuál es la etapa más importante?

# ¿ Minería de datos == BI?

- Múltiples conceptos: *data mining*, *business intelligence*, *data science*,*data analysis*, *feature engineering*
- BI es business intelligence o inteligencia de negocios
- Data science es ciencia de datos
- Analíitica avanzada
- Aprendizaje de máquina o Machine Learning
- otros...

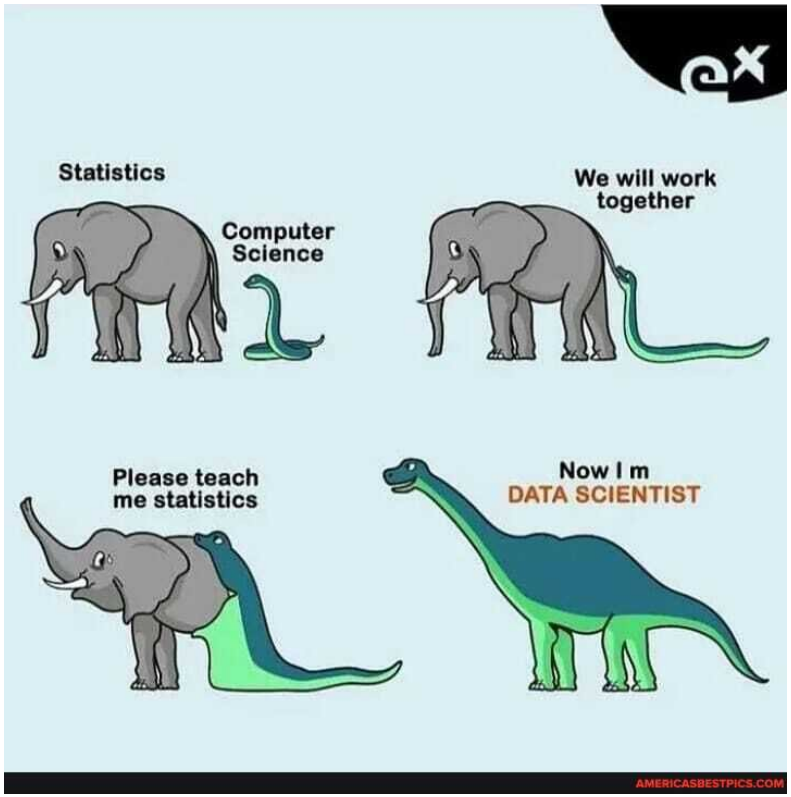
Muchos, muchos nombres, pero solo un origen. ¿Cuál es?

# Base transversal




- la Estadística y sus métodos son el componente principal de la disciplina.
- Machine learning está basado más allá, aprovechando los avances computacionales.
- Inteligencia Artificial aprovecha los conceptos de machine learning (y algunos más) para su desarrollo.

# Base transversal

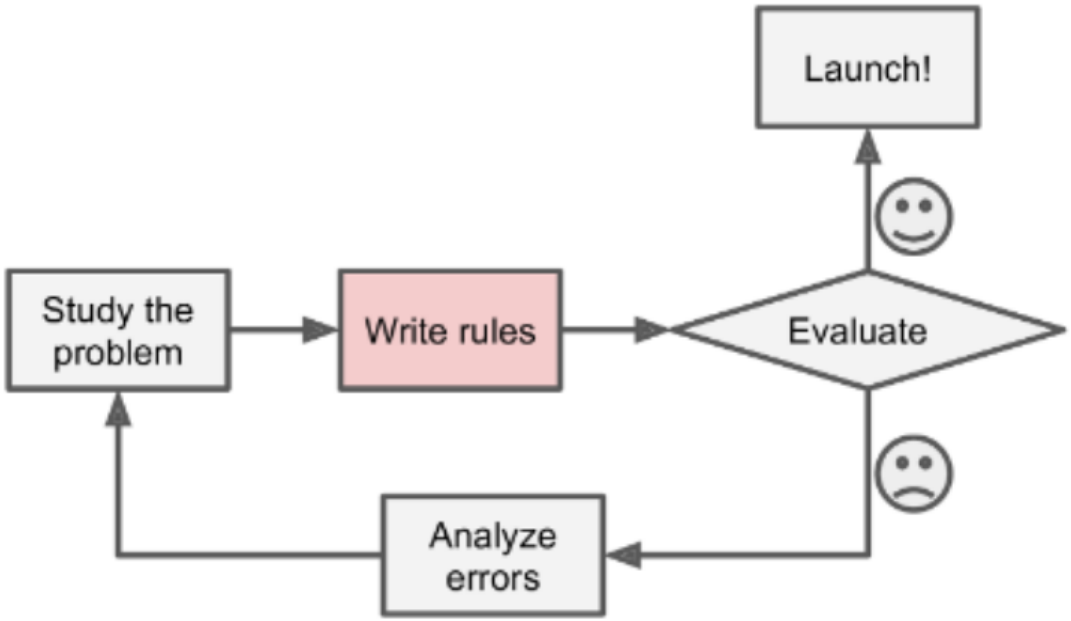



- Luego, si bien estadística es la fuente, ésta está concentrada en la creación de métodos,
- Machine learning está más concentrado en el uso de estos métodos y algoritmos.

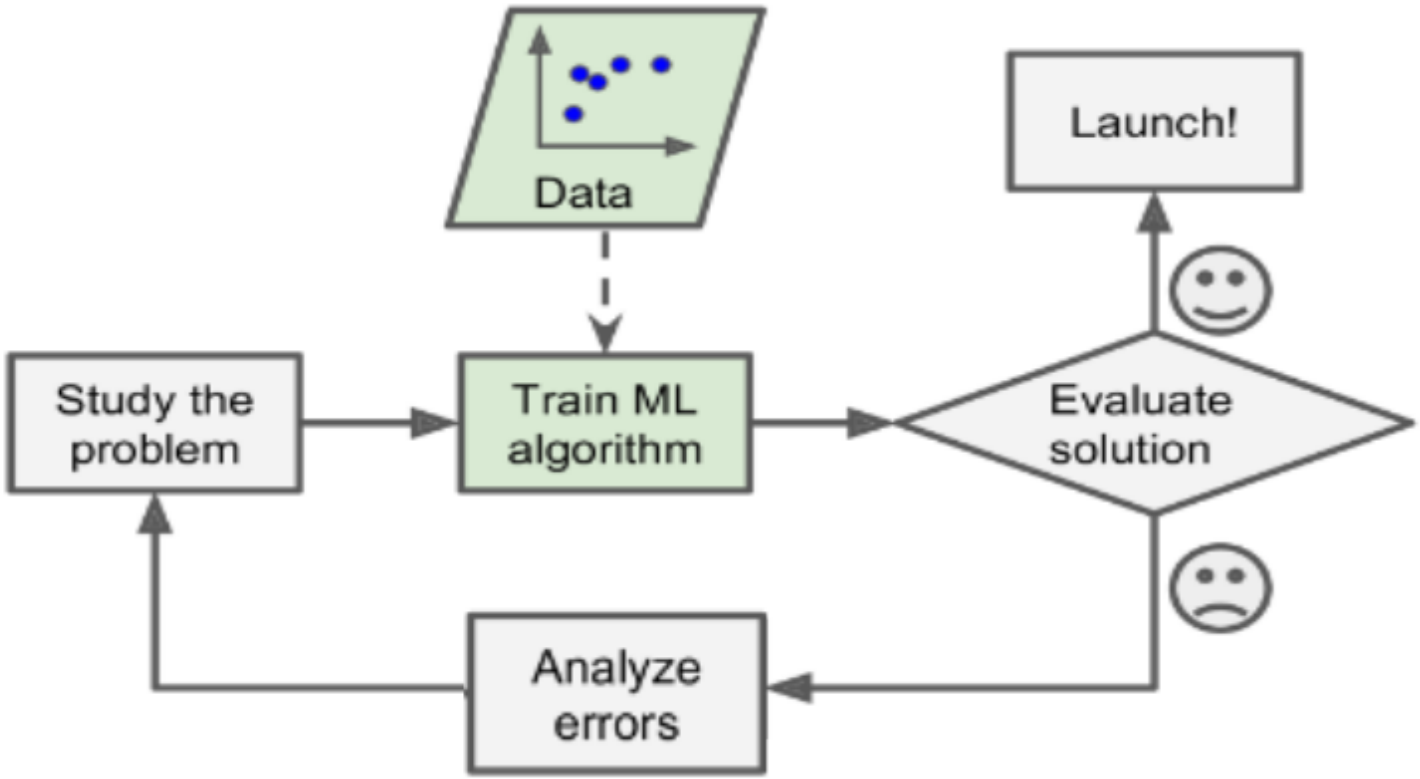
# ML en el día a día

- la carpeta de SPAM en el correo electrónico.
- predictor de palabras en whatsapp.
- predictor de búsqueda en google.

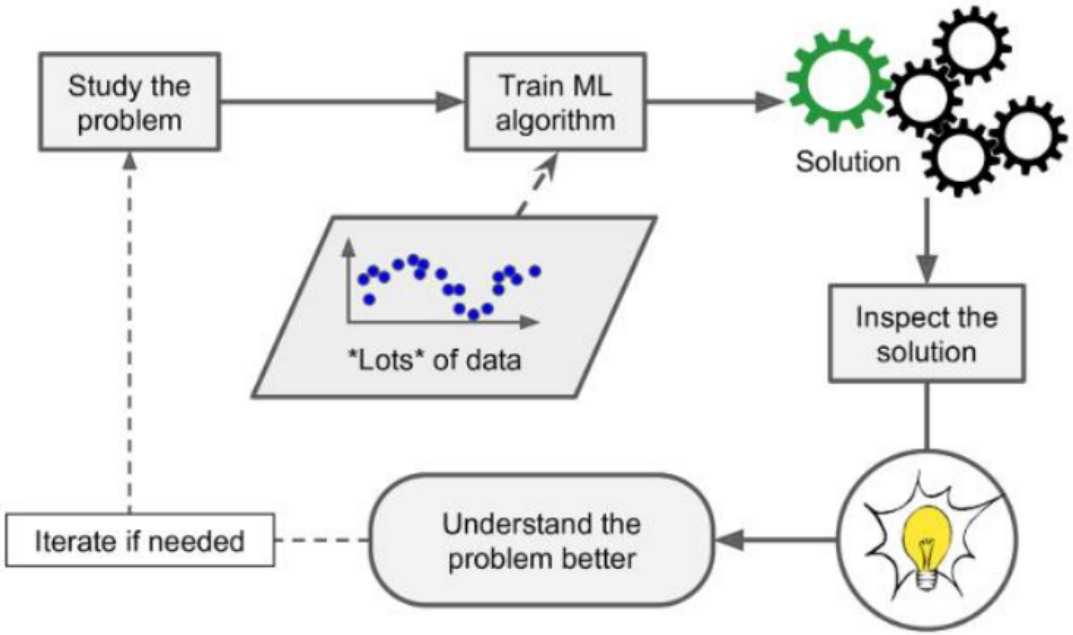
# ¿Cómo funciona?



# ¿Cómo funciona?



# ¿Cómo funciona?



# En Resumen

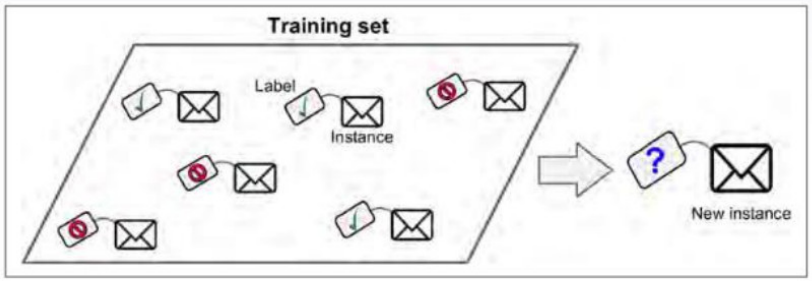
- Problemas para los que las soluciones existentes requieren muchos ajustes manuales o listas largas de reglas.
- Problemas complejos para los que no hay buenos resultados a partir de métodos tradicionales
- Entornos fluctuantes: Un sistema de aprendizaje automático puede adaptarse a nuevos datos
- Obtener conocimientos sobre problemas *complejos* y grandes cantidades de datos

# Tipos de sistemas de Machine Learning

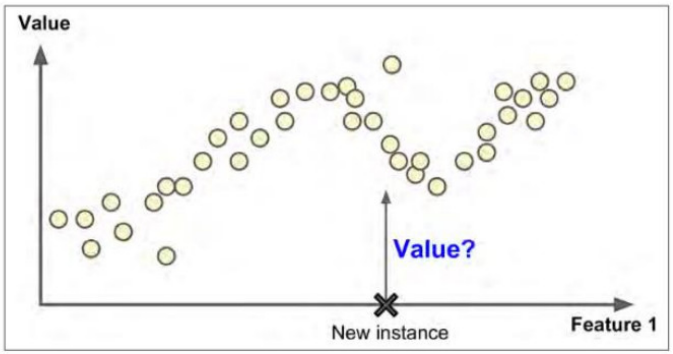
Aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje reforzado.

# Aprendizaje Supervisado

El objetivo de este aprendizaje es predecir el *valor* del ***output*** usando un conjunto de ***inputs***.
Tradicionalmente a los inputs se les llama *variables predictoras* y al output *variable respuesta* o *dependiente*




***Clasificación***



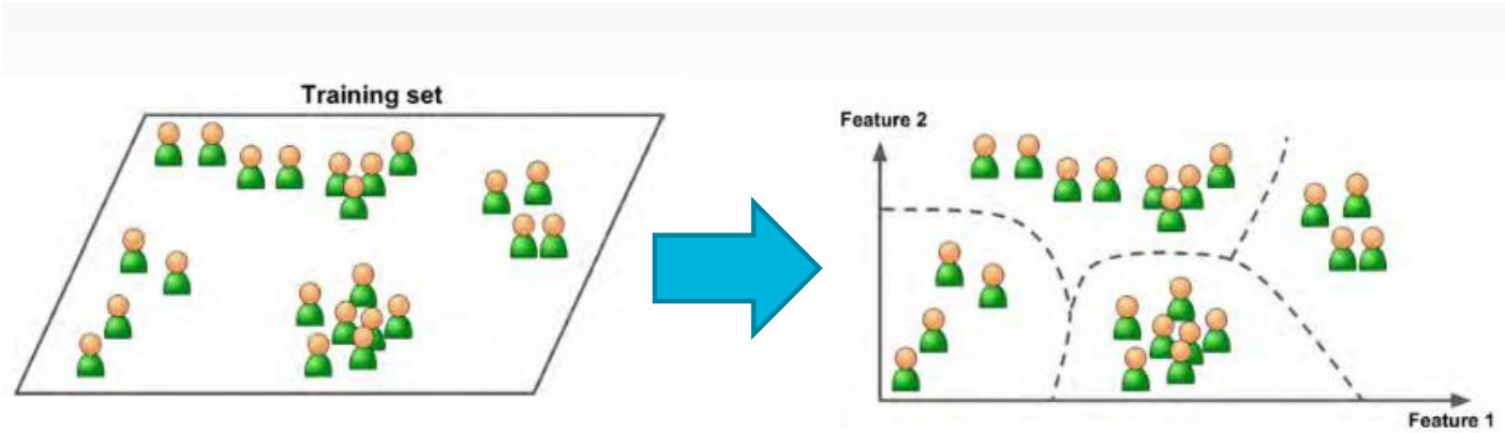

***Regresión***

*Ejemplos*

- Probabilidad de caer en incumplimiento
- Predicción de accidentes de un auto.

# Aprendizaje No Supervisado

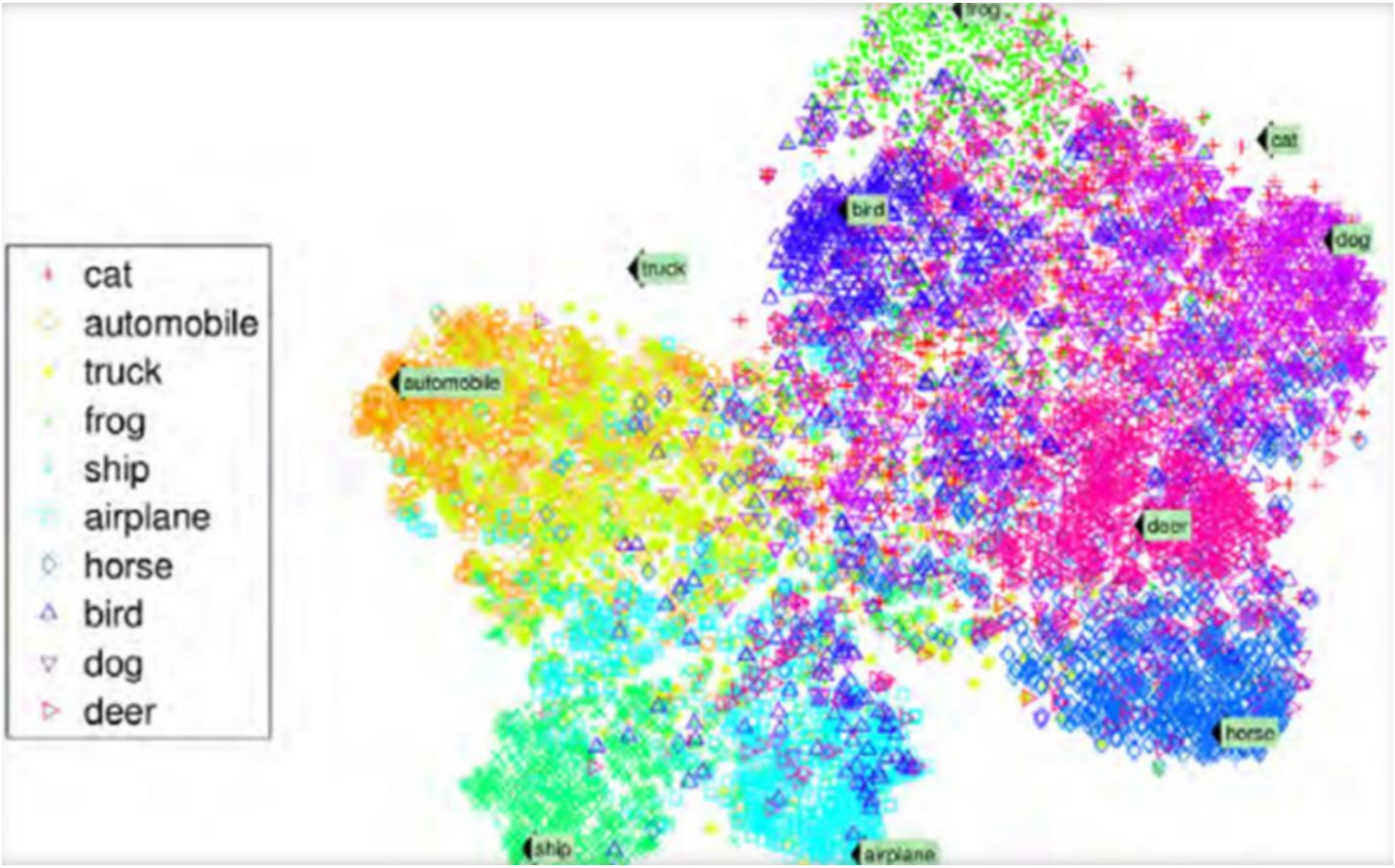
El objetivo de este aprendizaje es determinar la asociación entre variables predictoras cuando no se dispone de un ***output*** con el cual medir la precisión de las salidas.
En otras palabras, usualmente intentamos medir cuál es la distribución marginal de $\mathbb{P}(Y|X)$. En este caso, el objetivo es medir las propiedades de la densidad de probabilidad $\mathbb{P}(X)$ sin la ayuda de un $Y$.




*Ejemplos*

- Segmentación de una cartera de clientes

# Aprendizaje no supervisado

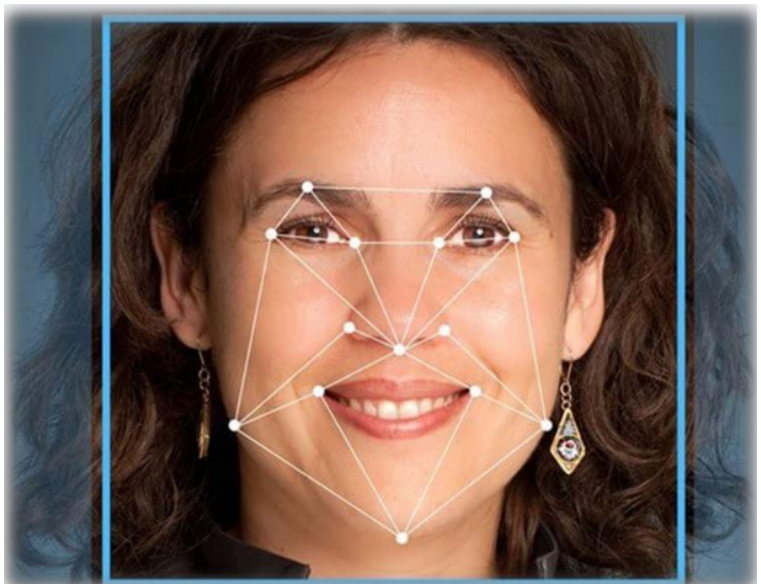


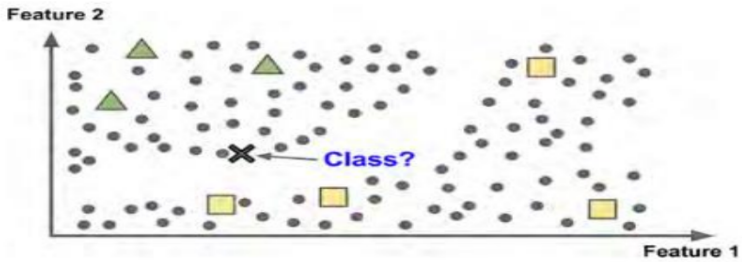
# Aplicaciones:
# Detección de anomalías

el sistema está entrenado con instancias ***normales*** y, cuando ve una nueva instancia, puede decir si se ve como una norml o si es probable que sea una anomalía.

# Aprendizaje semisupervisado

Google Photos reconoce automáticamente que la misma persona A aparece en las fotos 1, 5 y 11, mientras que otra persona B aparece en las fotos 2, 5 y 7. Esta es la parte no supervisada del algoritmo (**agrupamiento**). Ahora todo lo que necesita el sistema es que le digas quiénes son estas personas. Solo una etiqueta por persona y puede nombrar a todos en cada foto, lo cual es útil para hacer búsquedas.





# Aprendizaje reforzado

- ¿Qué pasa si no tenemos datos?
- ¿Podremos crear un sistema conductual que permita ver cuál será el comportamiento que tendrá el sistema?
- Esto es lo que hace el aprendizaje reforzado. Estableciendo parámetros iniciales, se "deja andar" el algoritmo y luego se revisan los resultados.