

Introducción a Machine Learning

Universidad Autónoma de Manizales
Semillero de Bioinformática e Inteligencia Artificial



Acreditación Institucional
DE ALTA CALIDAD
Resolución 009527 Mineducación Sep. 6 de 2019



Contenido

¿Qué es Machine Learning (ML)?

¿Por qué usar ML?

Tipos de sistemas de ML

Principales retos

¿Qué es Machine Learning (ML)?

Machine Learning (ML) es el área de estudio que da a los computadores la capacidad de aprender sin ser explícitamente programados.

(Arthur Samuel, 1959)

Desde el punto de vista de la ingeniería:

Un programa de computador se dice que aprende de la experiencia E con respecto a una tarea T y una medida de desempeño P , si su desempeño en T , medido por P , mejora con la experiencia E .

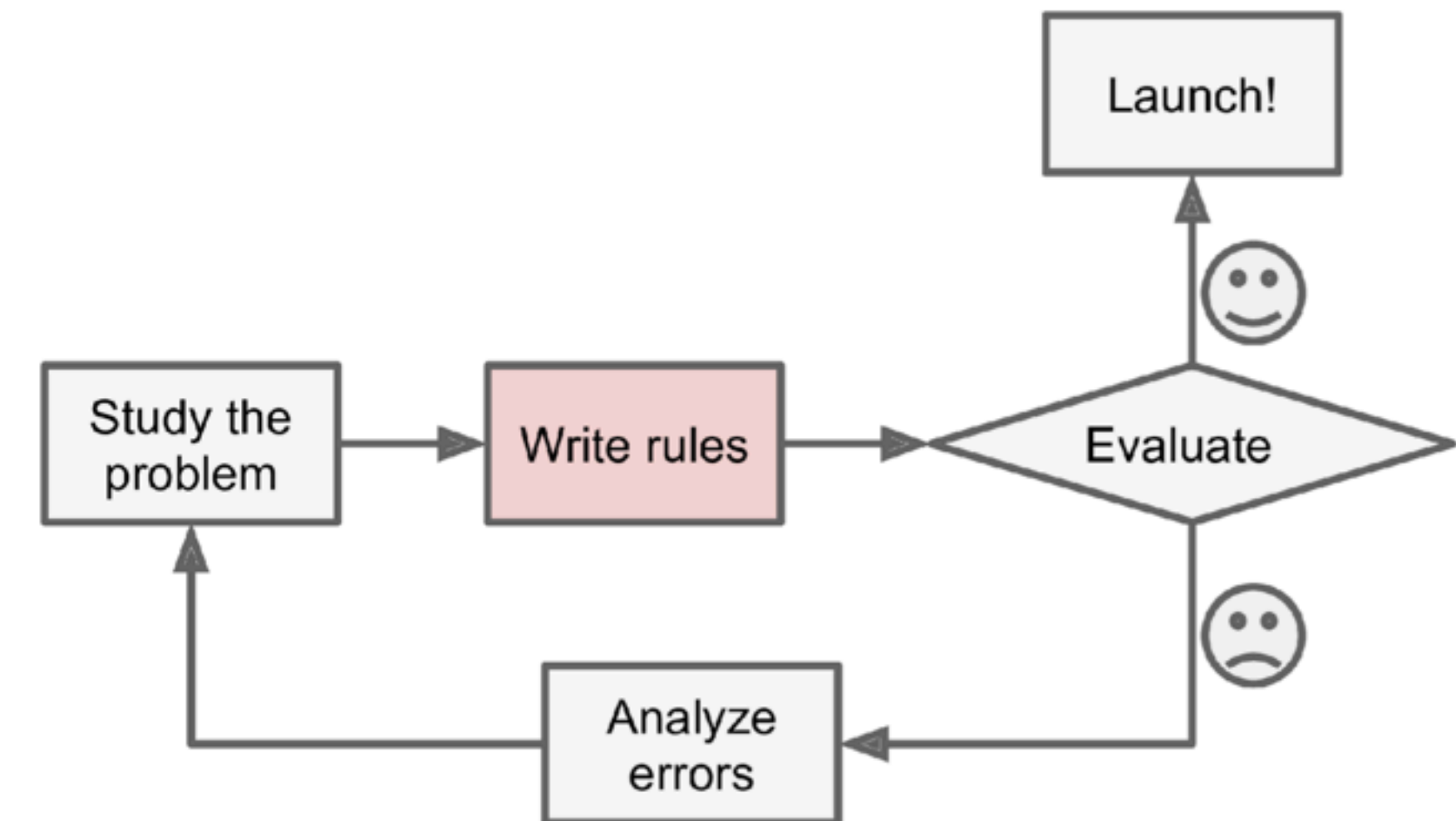
(Tom Mitchell, 1997)

¿Por qué usar ML?

Consideremos la aplicación de un filtro de spam en el correo electrónico:

¿Cómo escribir un filtro de spam usando técnicas de programación tradicionales?

1. Identificar las palabras o frases características de correos spam.
2. Identificar patrones en el remitente o cuerpo del correo.
3. Escribir un programa que detecte los rasgos identificados.

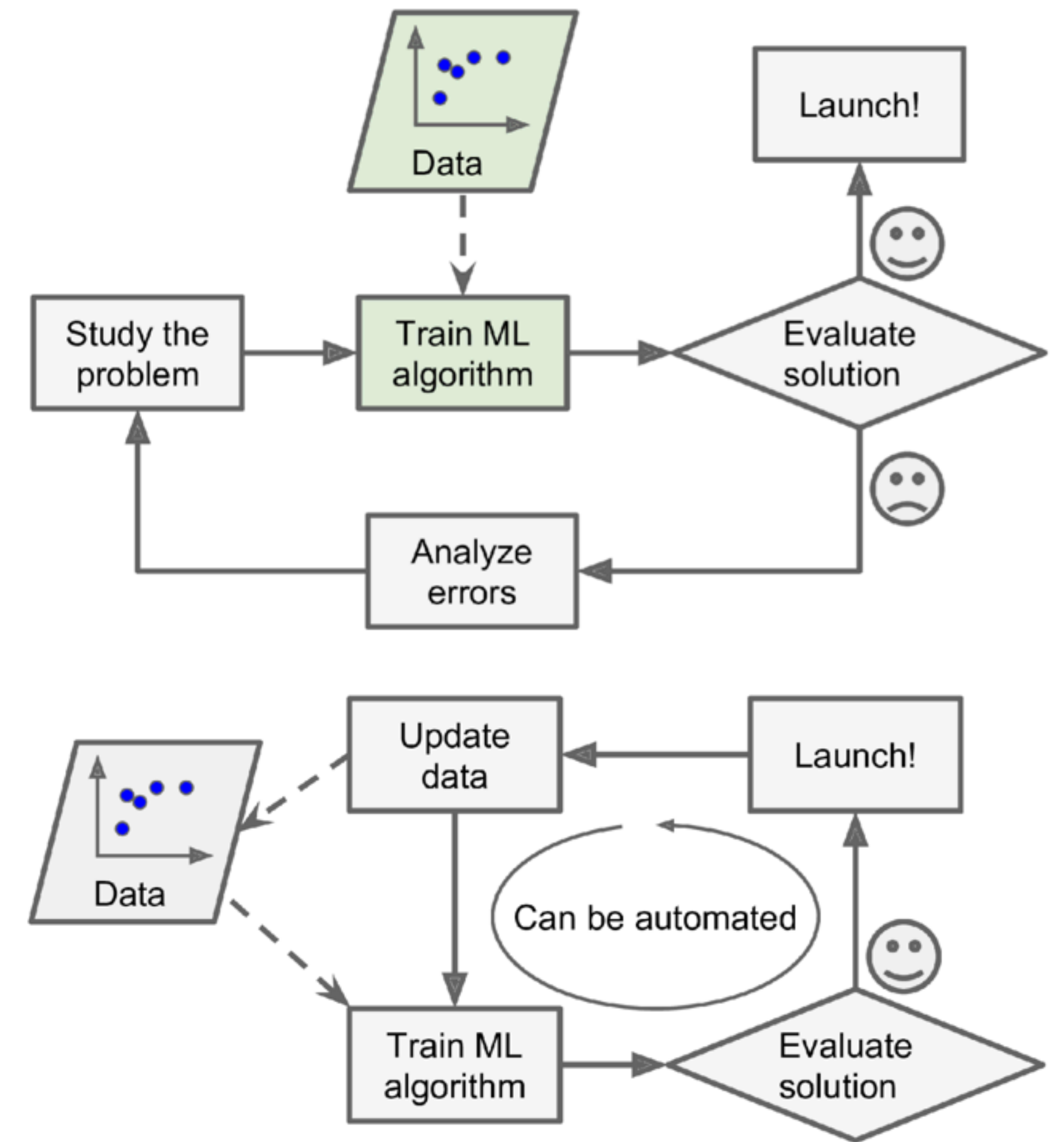


Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

¿Por qué usar ML?

En cambio, un filtro de spam basado en ML aprende automáticamente que palabras o frases son buenos predictores de un correo spam.

Puede ajustarse automáticamente a cambios en las palabras usadas por los spammers, sin necesidad de la intervenir o escribir nuevas reglas.



Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

¿Por qué usar ML?

También está el caso de tareas más complejas para las que definir un algoritmo o un conjunto de reglas se hace más complicado, por ejemplo:

- Reconocimiento del habla (Speech Recognition): se podría trabajar en el dominio de la frecuencia, pero considerando múltiples acentos e idiomas se vuelve una tarea muy compleja.
- Visión por computador (Computer Vision): los modelos de ML han sobrepasado el desempeño de algoritmos tradicionales basados en procesamiento digital de imágenes.

Tipos de sistemas de ML

Si se entrenan o no con supervisión humana:

- Aprendizaje supervisado
- Aprendizaje semi-supervisado
- Aprendizaje no supervisado
- Aprendizaje por refuerzo

Si tienen la capacidad de aprender a medida que se generan nuevos datos (Aprendizaje online vs por lotes).

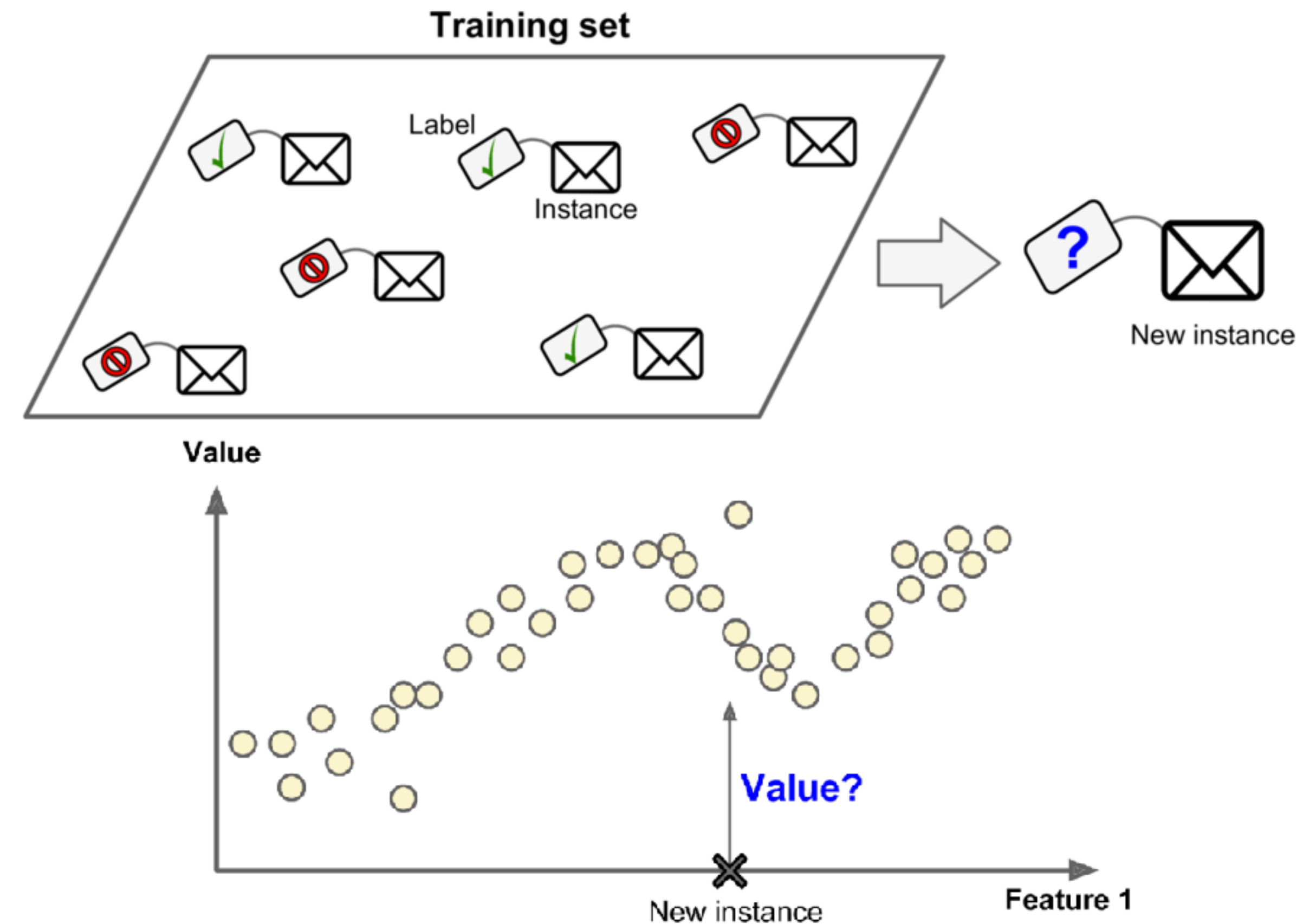
Si funcionan simplemente comparando nuevas muestras a muestras conocidas, o generan un modelo predictivo en base a las muestras conocidas (Aprendizaje basado en instancias vs basado en modelos).

Tipos de sistemas de ML

Aprendizaje supervisado

El conjunto de entrenamiento, además de las características (*features*) o atributos, contiene las soluciones u objetivos (*etiquetas/labels*). Las tareas típicas son la clasificación (las etiquetas son categorías) y la regresión (las etiquetas son valores numéricos).

Incluye algoritmos como *k-Nearest Neighbors*, *Linear Regression*, *Logistic Regression*, *Support Vector Machines*, *Decision Trees* y *Neural Networks*.



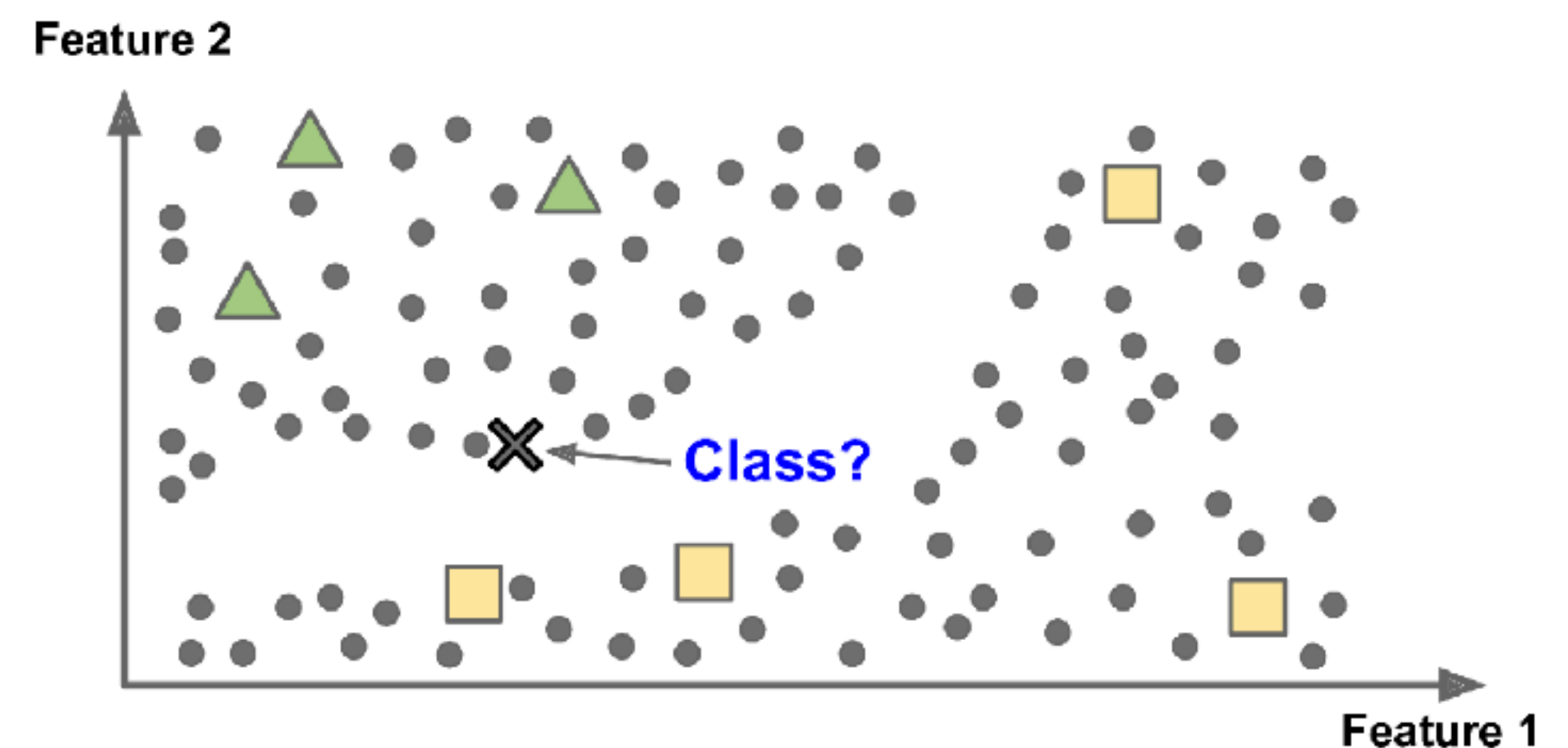
Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tipos de sistemas de ML

Aprendizaje semi-supervisado

En muchas ocasiones, la tarea de etiquetar los datos es costosa y requiere de mucho tiempo, una alternativa es etiquetar solo unas cuantas muestras y trabajar con datasets parcialmente etiquetados.

Por ejemplo, el reconocimiento de personas en fotos por Google Photos.



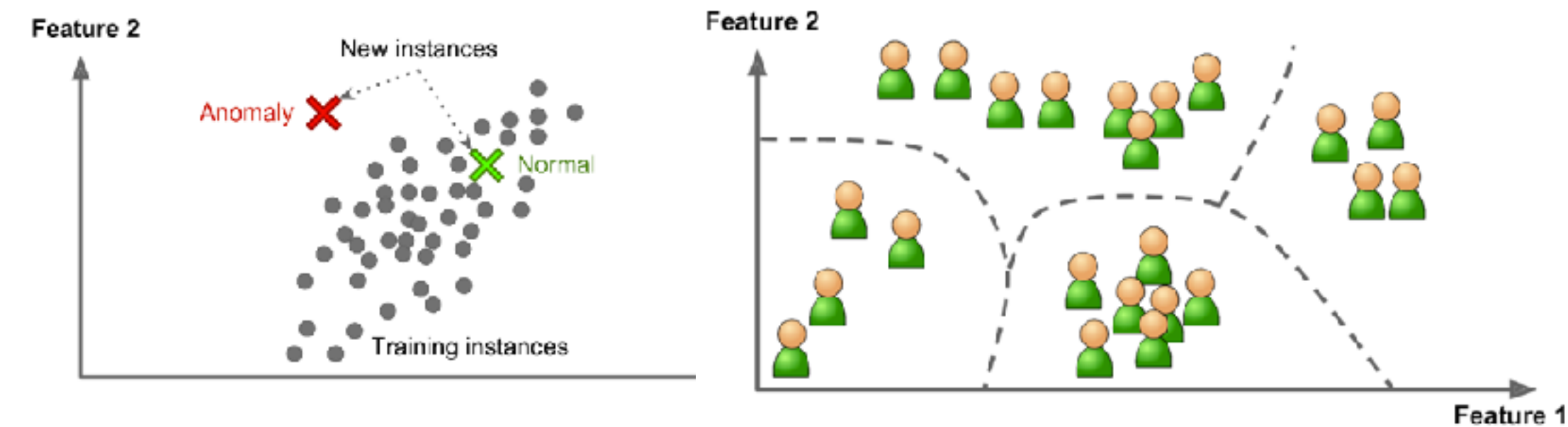
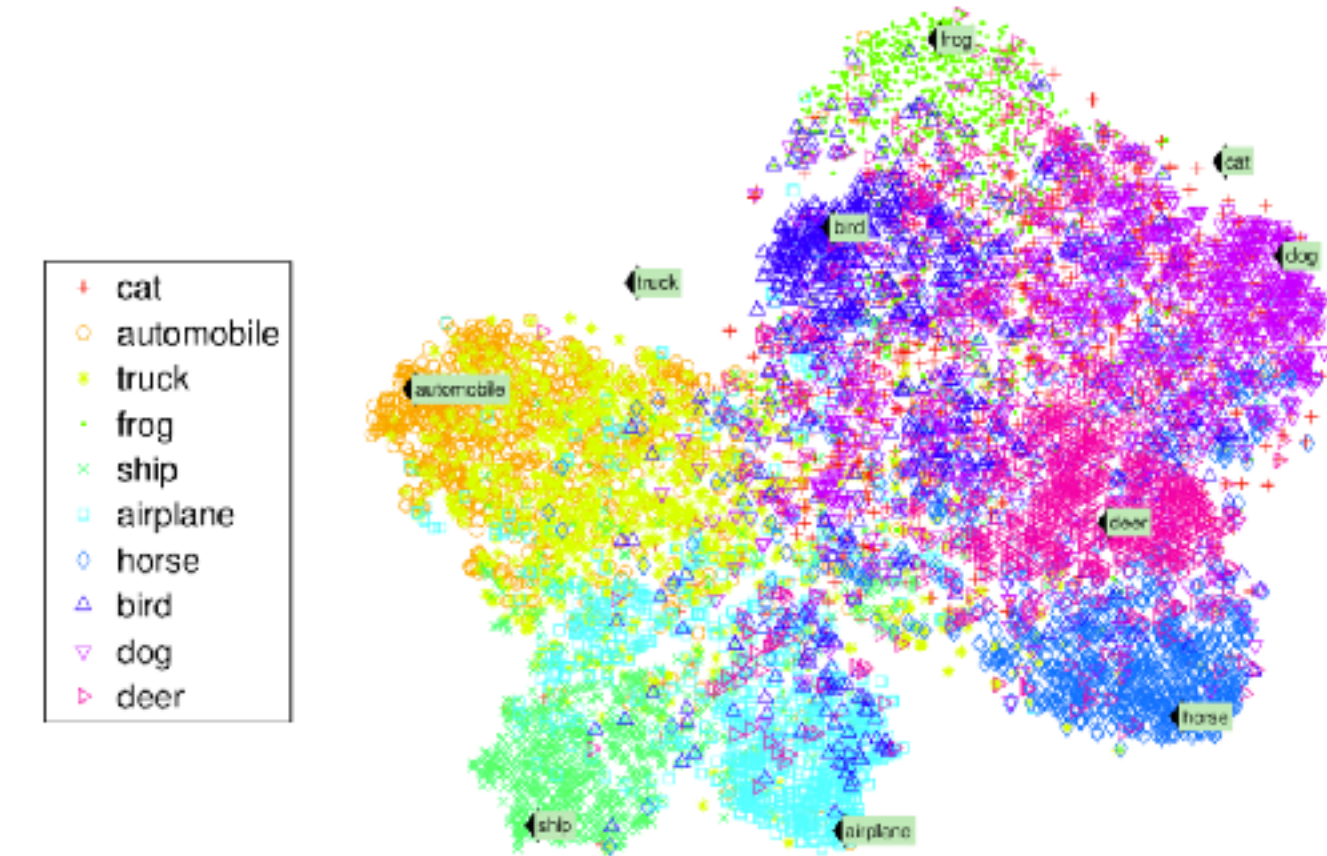
Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tipos de sistemas de ML

Aprendizaje no supervisado

Este tipo de aprendizaje trabaja con datasets no etiquetados. Las tareas más importantes son:

- Agrupamiento (*Clustering*)
 - K-Means
 - Hierarchical Cluster Analysis (HCA)
- Detección de anomalías
 - One-class SVM
- Visualización y reducción de dimensionalidad
 - Principal Component Analysis (PCA)
 - t-Distributed Stochastic Neighbor Embedding (t-SNE)

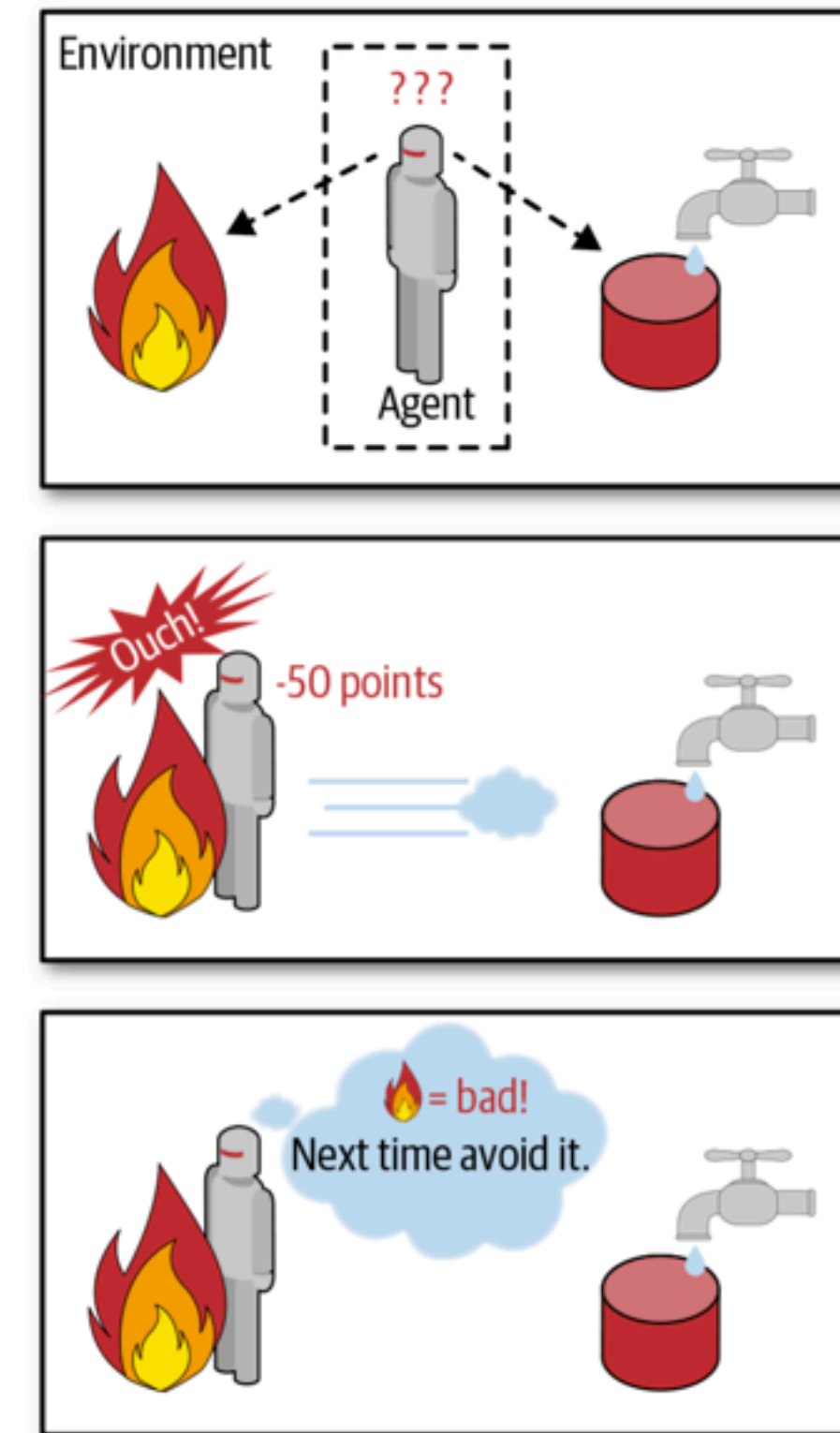


Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tipos de sistemas de ML

Aprendizaje por refuerzo

El sistema de aprendizaje, llamado *agente* en este contexto, selecciona y ejecuta acciones, por las cuales recibe penalizaciones o recompensas. El objetivo es que el agente desarrolle la estrategia o política que maximice las recompensas con el tiempo.



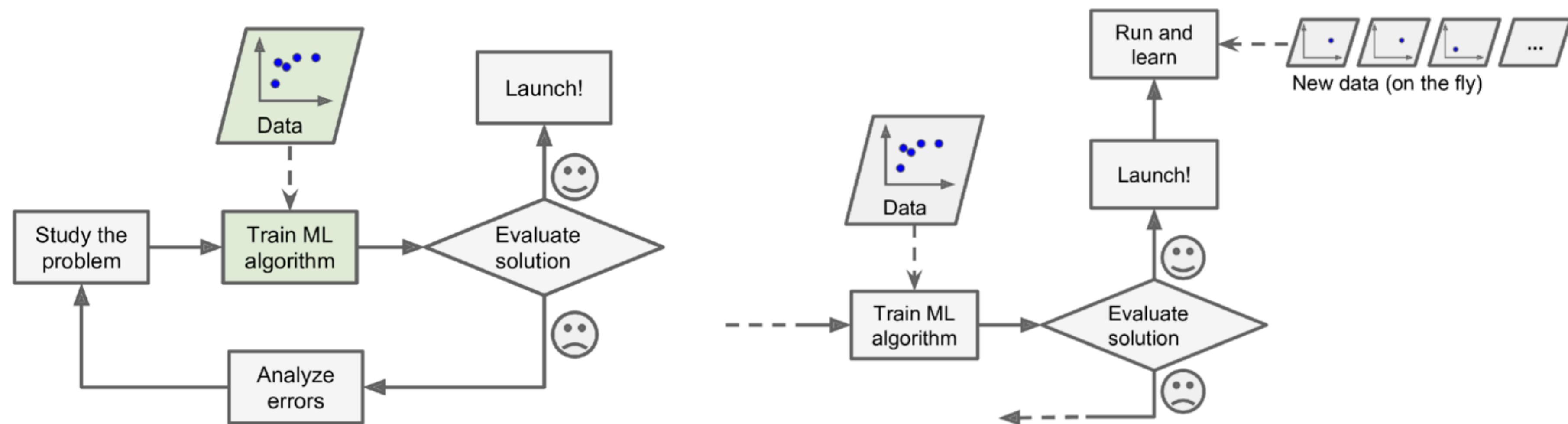
- 1 Observe
- 2 Select action using policy
- 3 Action!
- 4 Get reward or penalty
- 5 Update policy (learning step)
- 6 Iterate until an optimal policy is found

Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tipos de sistemas de ML

Aprendizaje online vs por lotes

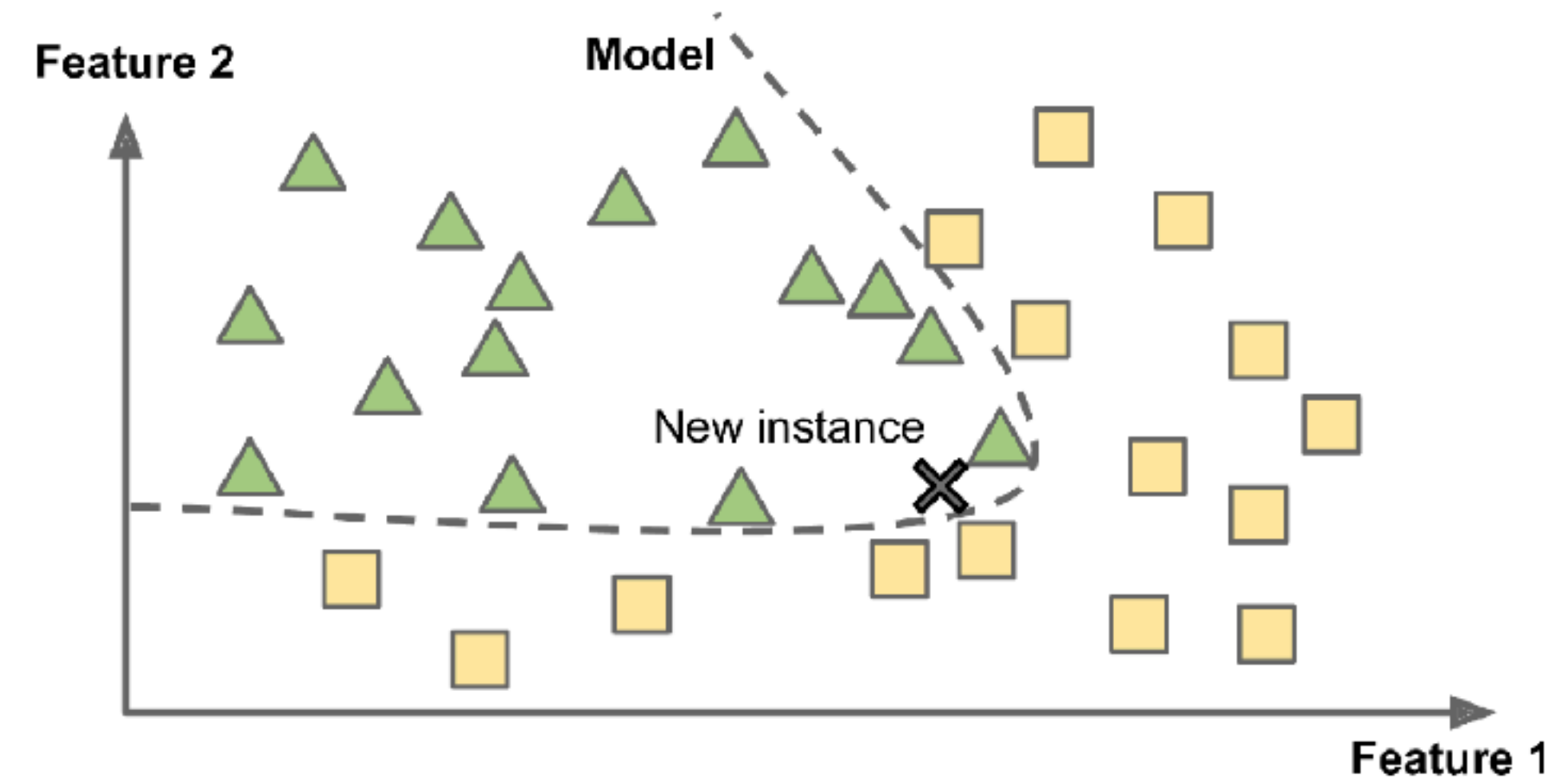
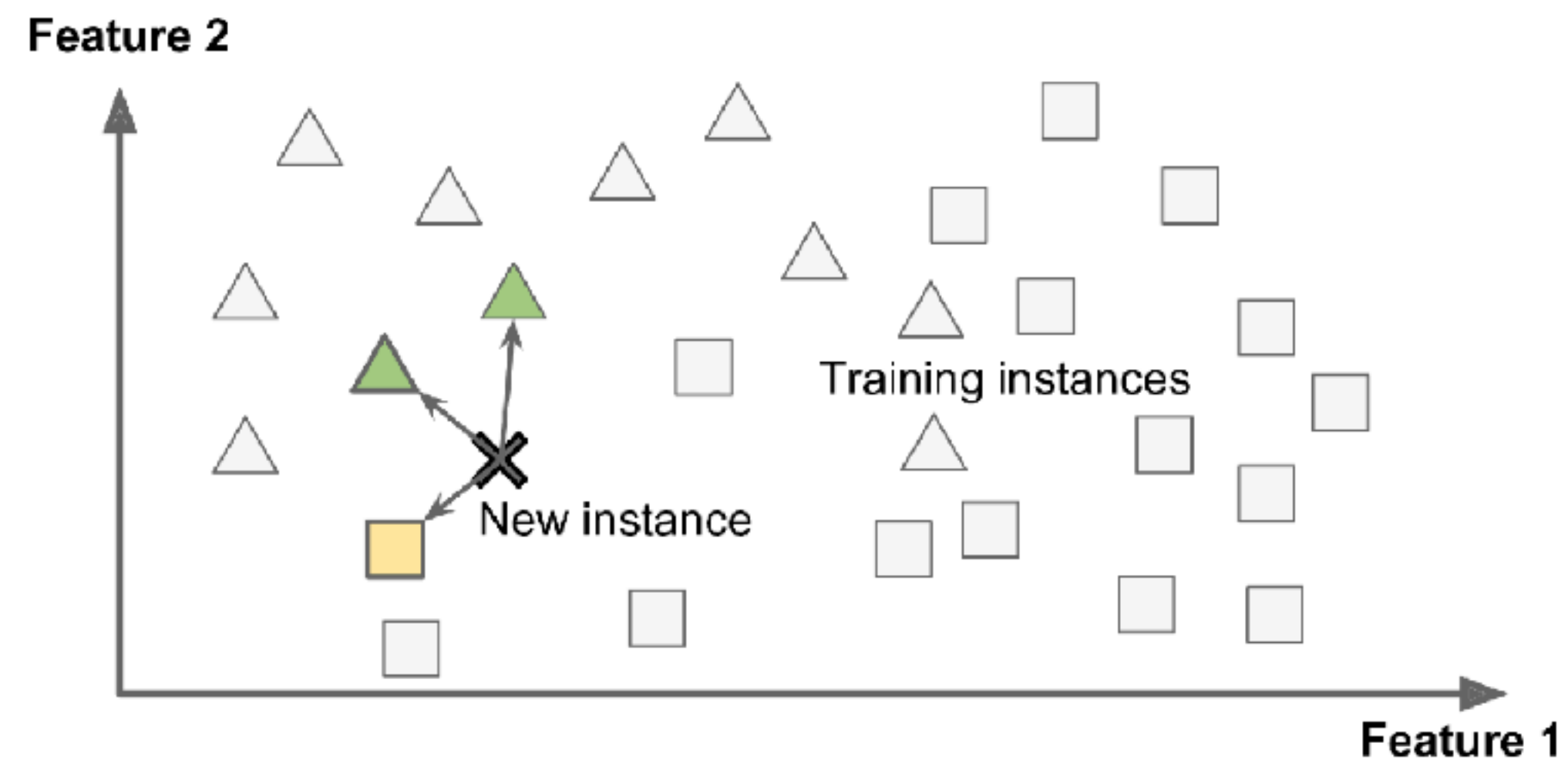
En el aprendizaje por lotes el sistema se entrena una vez *offline* y se despliega a producción; en el aprendizaje online el modelo tiene la capacidad de reentrenarse luego de estar en producción con los nuevos datos.



Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Tipos de sistemas de ML

Aprendizaje basado en instancias vs basado en modelos



Geron, A. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Principales retos

- Cantidad insuficiente de datos
 - En la mayoría de los casos, tener más datos implica mejores resultados
- Datos de entrenamiento no representativos (*Data mismatch*)
 - El objetivo final es tener un buen desempeño en nuevas muestras o datos no conocidos, por lo que se requieren datos de entrenamiento con características similares
- Datos de baja calidad
- Características irrelevantes
- Sobreajuste (*Overfitting*) de los datos de entrenamiento
- “Subajuste” (*Underfitting*) de los datos de entrenamiento

