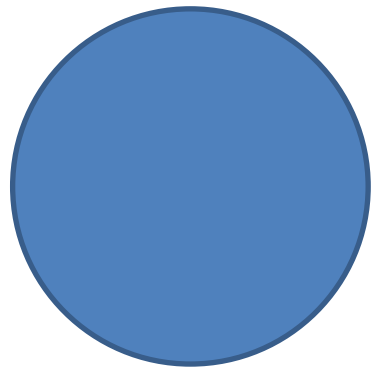


Machine Learning & Deep Learning

Introducción al Machine Learning

Profesor: Carlos Moreno Morera



Contenido

En este tema se introducen los conceptos básicos de Machine Learning y Deep Learning necesarios para poder entender los distintos algoritmos y técnicas empleadas en estas áreas.

01 ¿Qué es el ML?

02 Metodología del ML

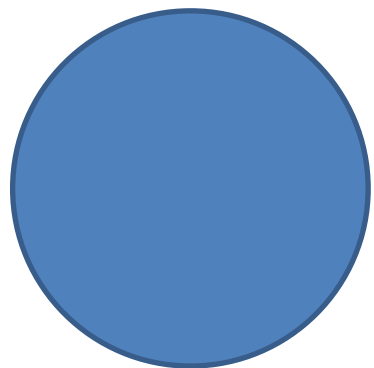
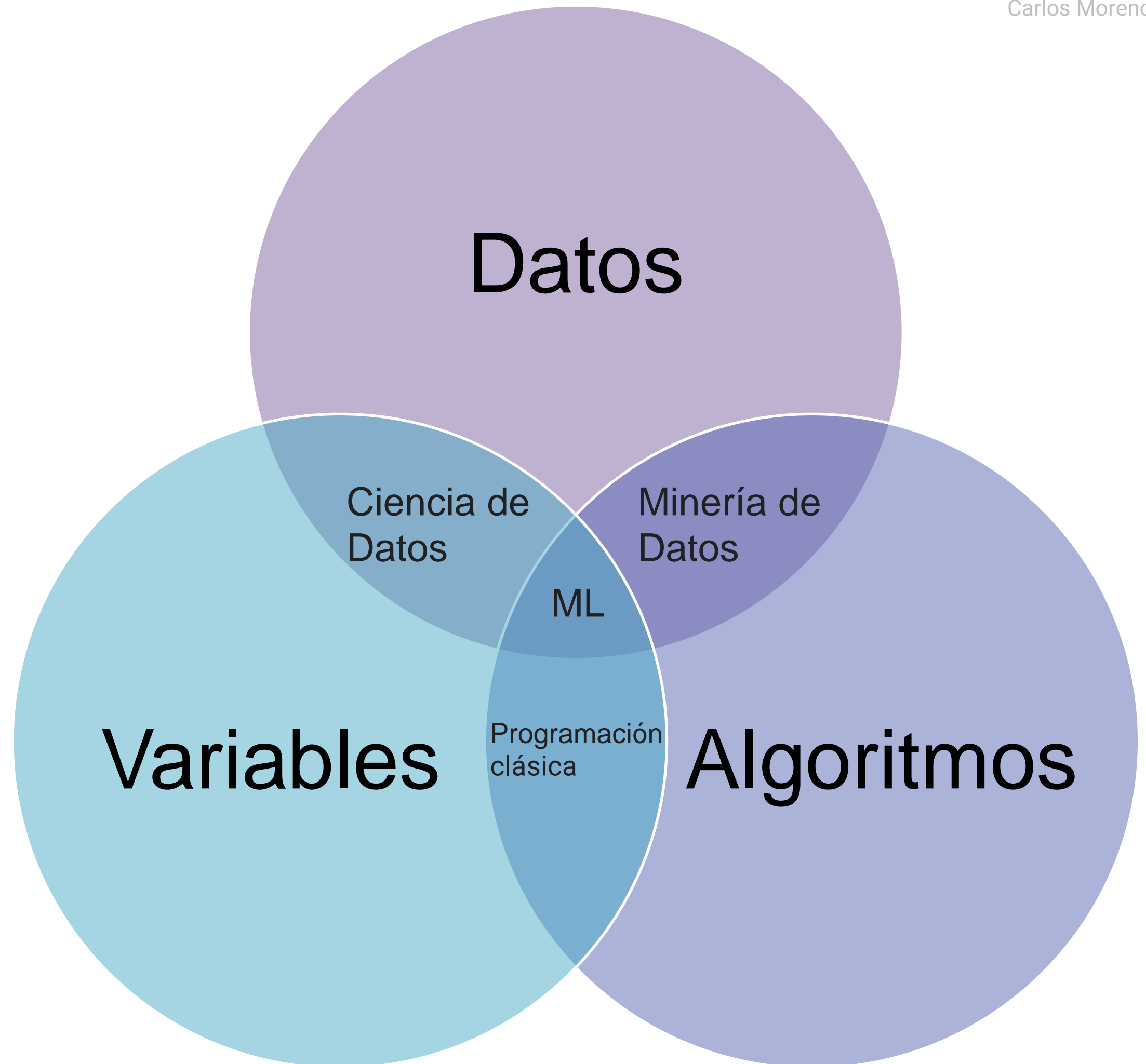
03 Técnicas de selección de datos

04 Scikit-Learn

01

¿Qué es el ML?

Conceptos básicos y componentes principales del Aprendizaje Automático.



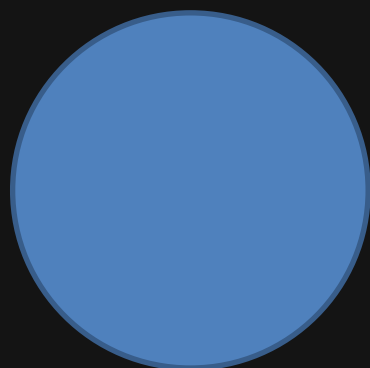
“

El "big data", como me gusta decir, es una gran cantidad de basura que esconde diamantes. Entonces lo que hacen los algoritmos de "machine learning" es revelar y mostrar cuáles son esos diamantes.Cuál es la información importante

”

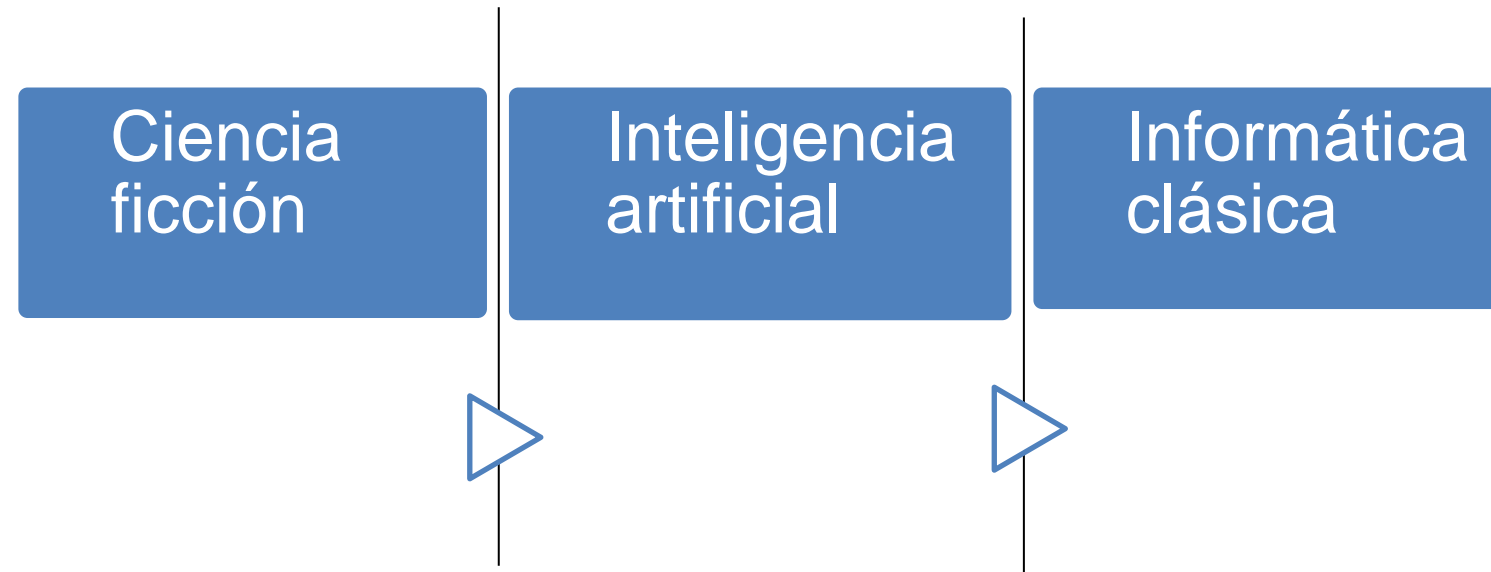
Todd Yellin

Vicepresidente de producto de Netflix

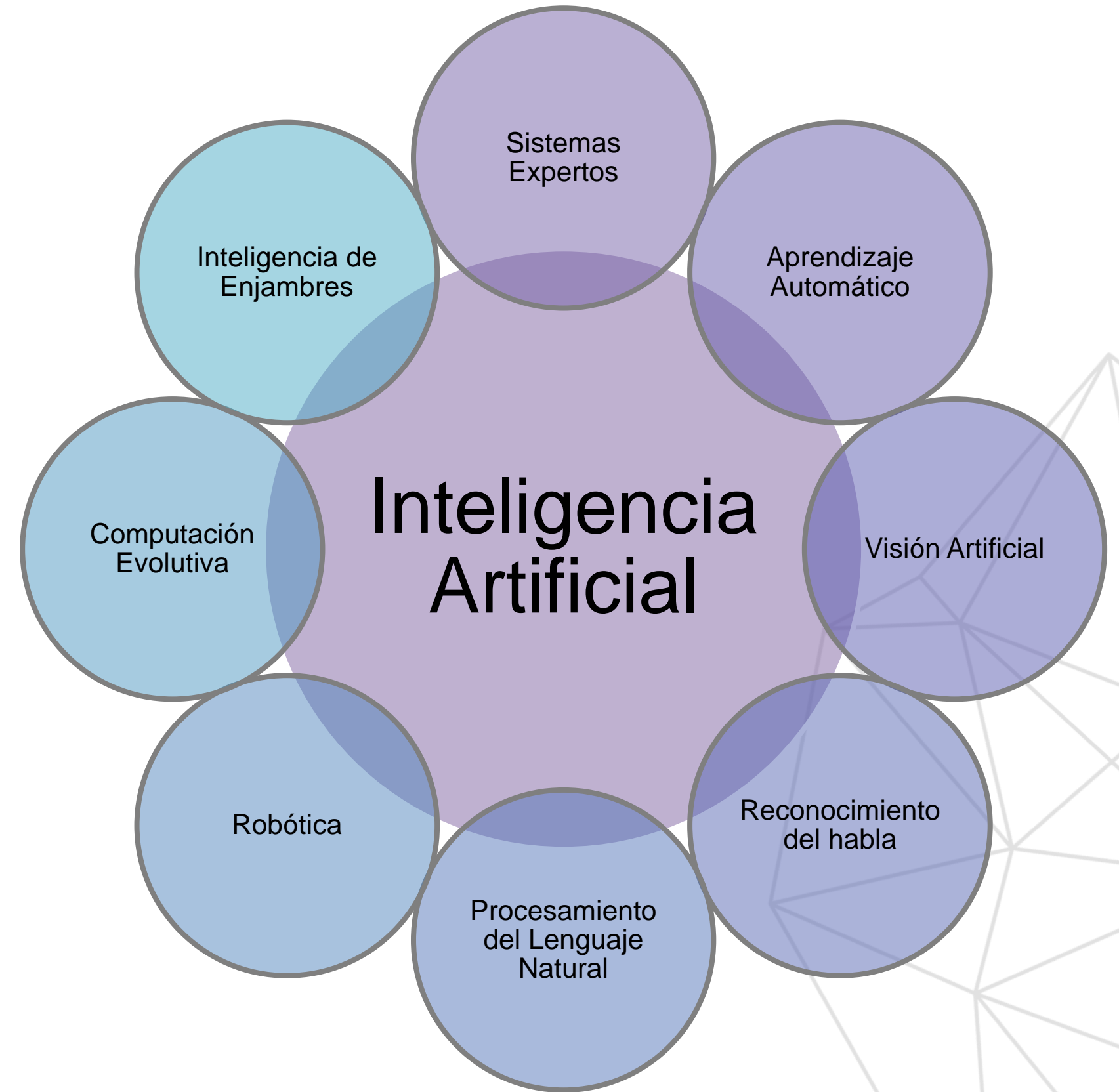


Conceptos clave: Inteligencia Artificial

- **Definición:** Disciplina que desarrolla metodologías para resolver problemas que no tienen fácil solución algorítmica.
- **Propósito:** Analizar los mecanismos que dan lugar a conductas inteligentes para reproducir dichas conductas en máquinas

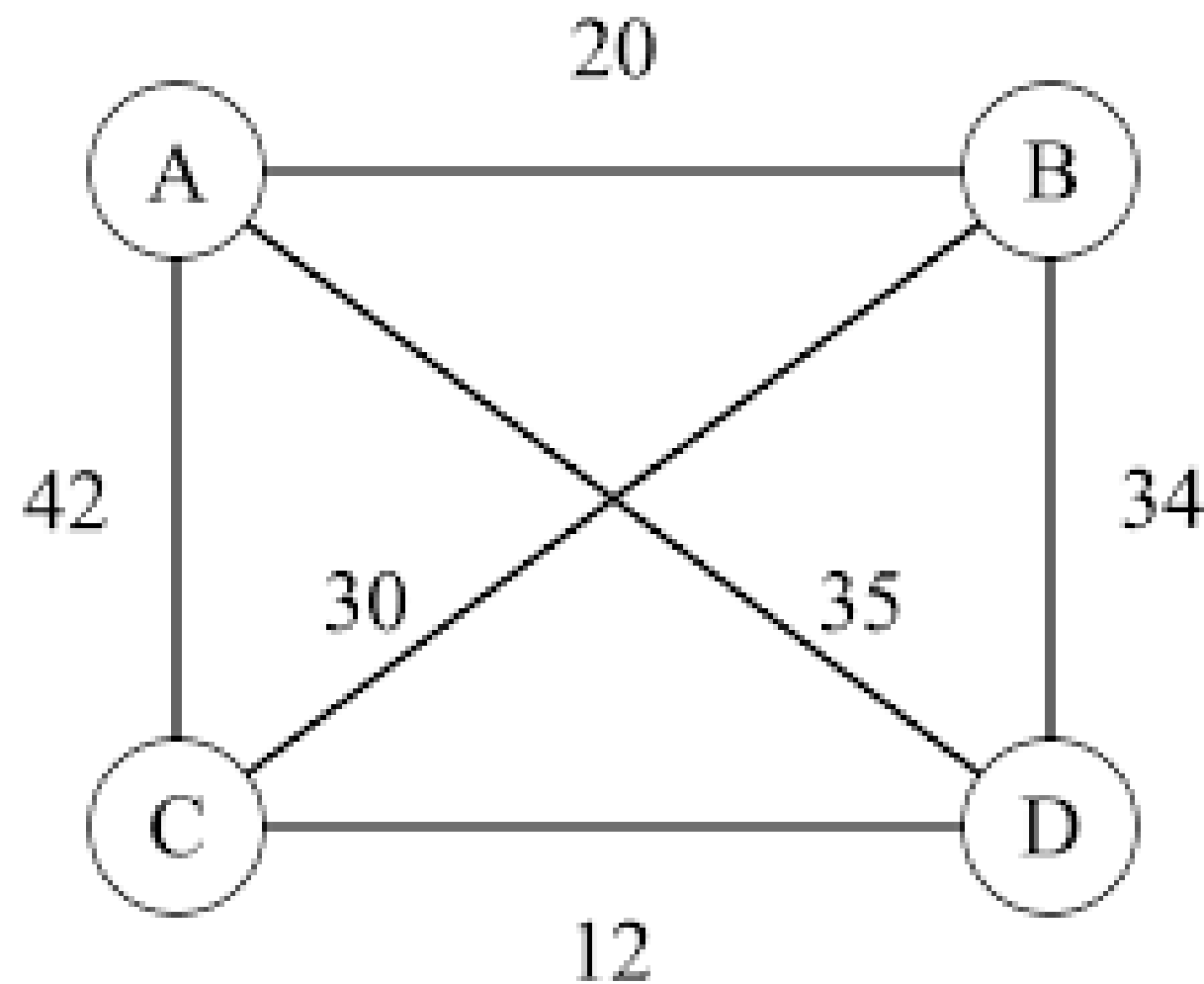


Frontera móvil de la IA



Conceptos clave: Inteligencia Artificial

Problema del viajante: Dada una lista de ciudades y las distancias entre cada par de ellas, ¿cuál es la ruta más corta posible que visita cada ciudad exactamente una vez y al finalizar regresa a la ciudad origen?

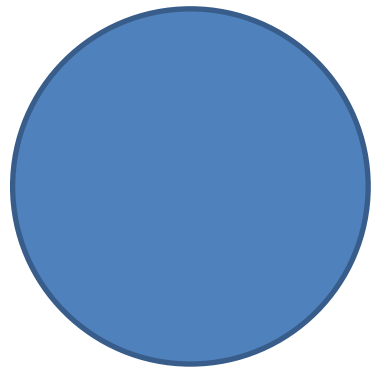


Con 30 ciudades:

- Más de $4 \cdot 10^{30}$ rutas posibles
- Calculando un millón de rutas por segundo tardaríamos...

¡ 10^{17} años!

La edad del universo es de 13.400 millones de años



Conceptos clave: Aprendizaje automático

- El Aprendizaje automático es la rama de la Inteligencia Artificial que busca cómo dotar a las máquinas de la capacidad de aprendizaje.

Arthur Samuel (1959)

- Un programa se dice que aprende de una experiencia E con respecto a alguna tarea T y alguna medida de rendimiento R , si su rendimiento en T medido con R , mejora con la experiencia E .

Tom Mitchell (1998)

Por lo tanto se entiende el **aprendizaje** como la **generalización del conocimiento a partir** de un conjunto de **experiencias**.



Conceptos clave: Aprendizaje automático

Programación Explícita

Reglas



Datos

Historial	Deudas	Avalista	Promesa
Malo	No	Si	Si

Programa



Resultado

Concedido

Conceptos clave: Aprendizaje automático

Datos

Historial	Deudas	Avalista	Promesa
Malo	No	Si	Si
...
Bueno	Si	No	Si

Aprendizaje Automático

Algoritmo de Aprendizaje



Resultados

Concesión
Concedido
...
No Concedido

Reglas | Modelo



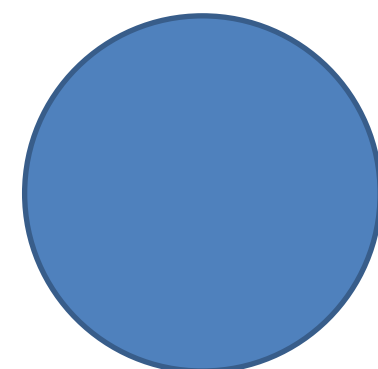
Conceptos clave: Aprendizaje automático

- El área del Machine Learning **tiene como objetivo entrenar modelos** (con Algoritmos de Aprendizaje) para que posteriormente sean capaces de realizar tareas de forma autónoma.
- Las técnicas utilizadas en Machine Learning, aparecen también en otras ramas de la Inteligencia Artificial, ya que el resto de capacidades inteligentes que se tratan de lograr pueden obtenerse ya sea mediante programación explícita o aprendizaje. Pero **no hay que confundir la parte por el todo**.

Observación

Precio	25.000\$	24.000\$	23.000\$	22.000\$
Edad	0	1	2	3

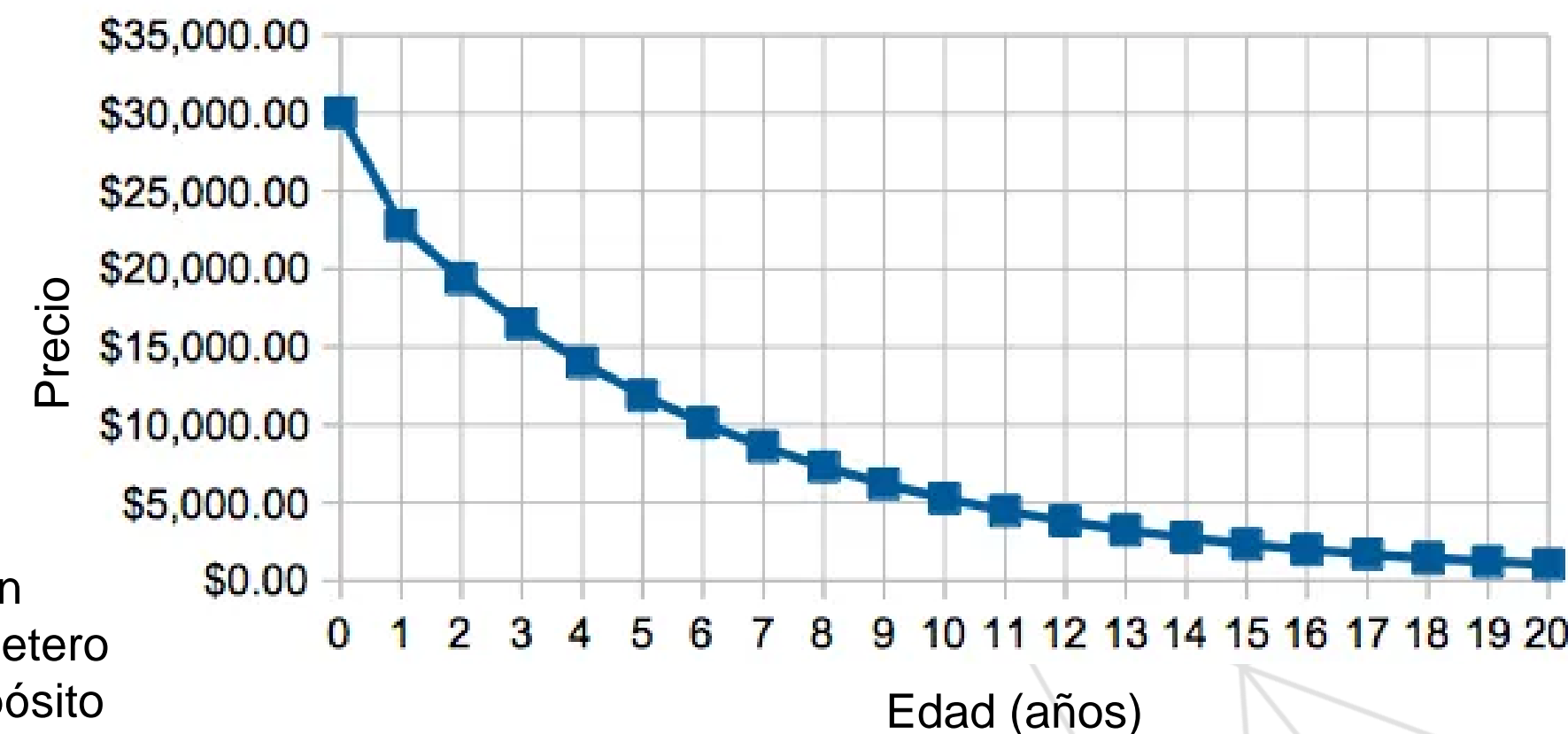
Dedución: El precio de los coches depende de su edad y se reduce en 1.000\$ con cada año



Realidad

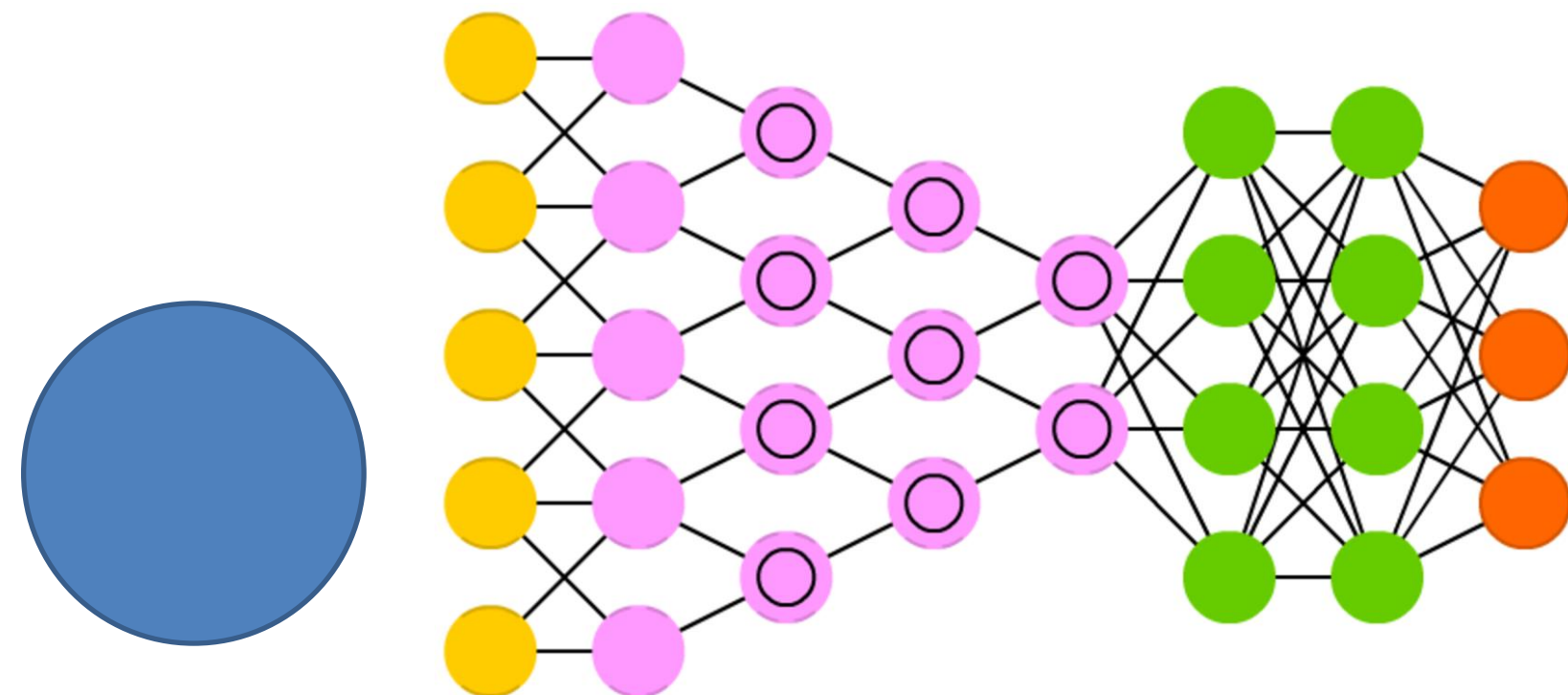
- Marca
- Modelo
- Edad
- Kilometraje
- Cilindrada
- Tipo de combustible
- Cambio
- Longitud
- Tamaño de neumáticos
- Número de plazas
- Potencia
- Tipo de tracción
- Capacidad maletero
- Capacidad depósito
- ... y muchas más

Depreciación de coches



Conceptos clave: Redes neuronales

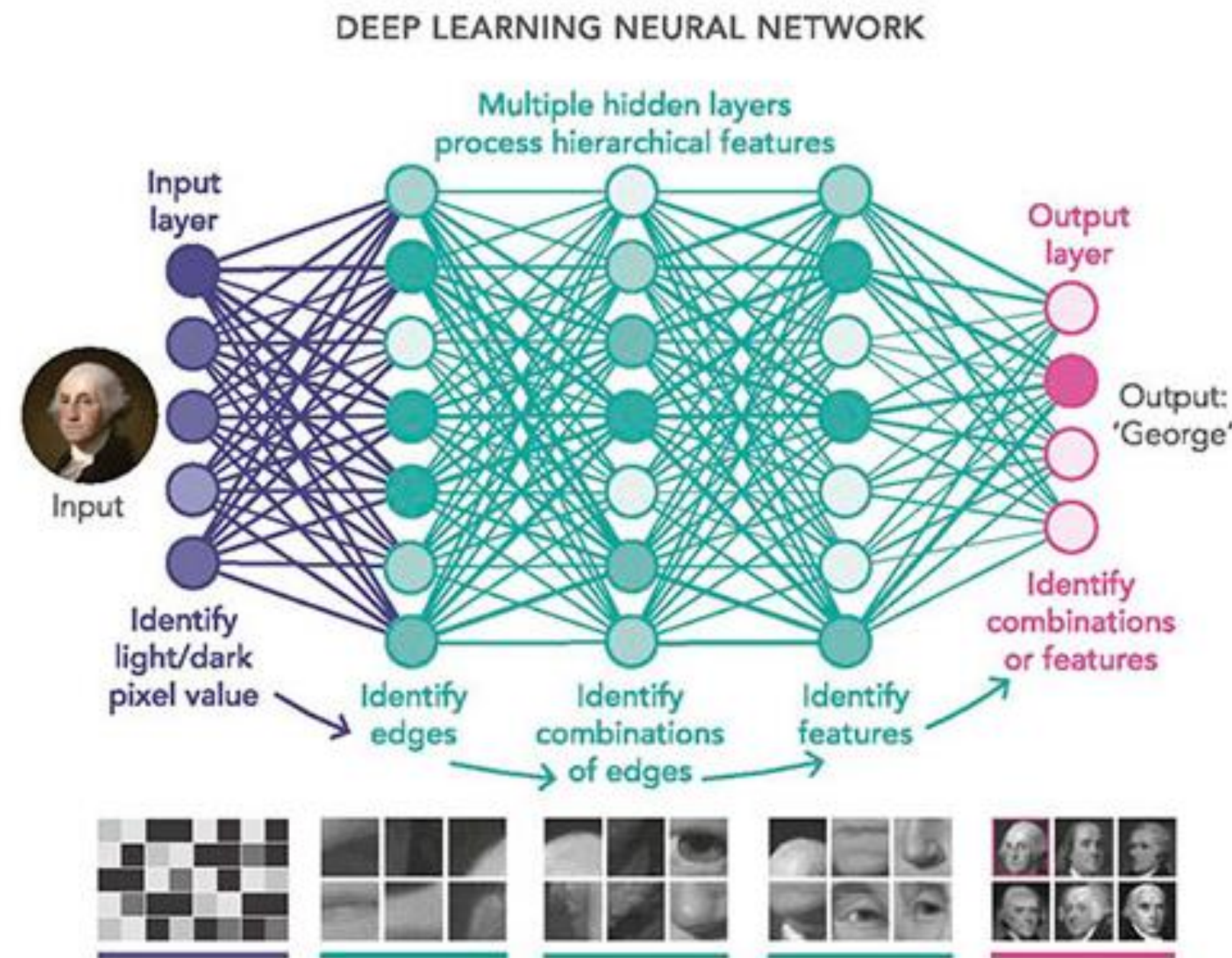
- Las redes neuronales artificiales (RNA, o ANN en inglés) son un tipo de solución dentro de las distintas técnicas del Machine Learning (nace en los años 50).
- Al igual que las neuronas humanas, las neuronas artificiales se interconectan entre sí con el fin de transmitirse información.
Aprenden de forma jerarquizada: de conceptos concretos a abstractos



Pero, ¿cuántas capas podemos poner?

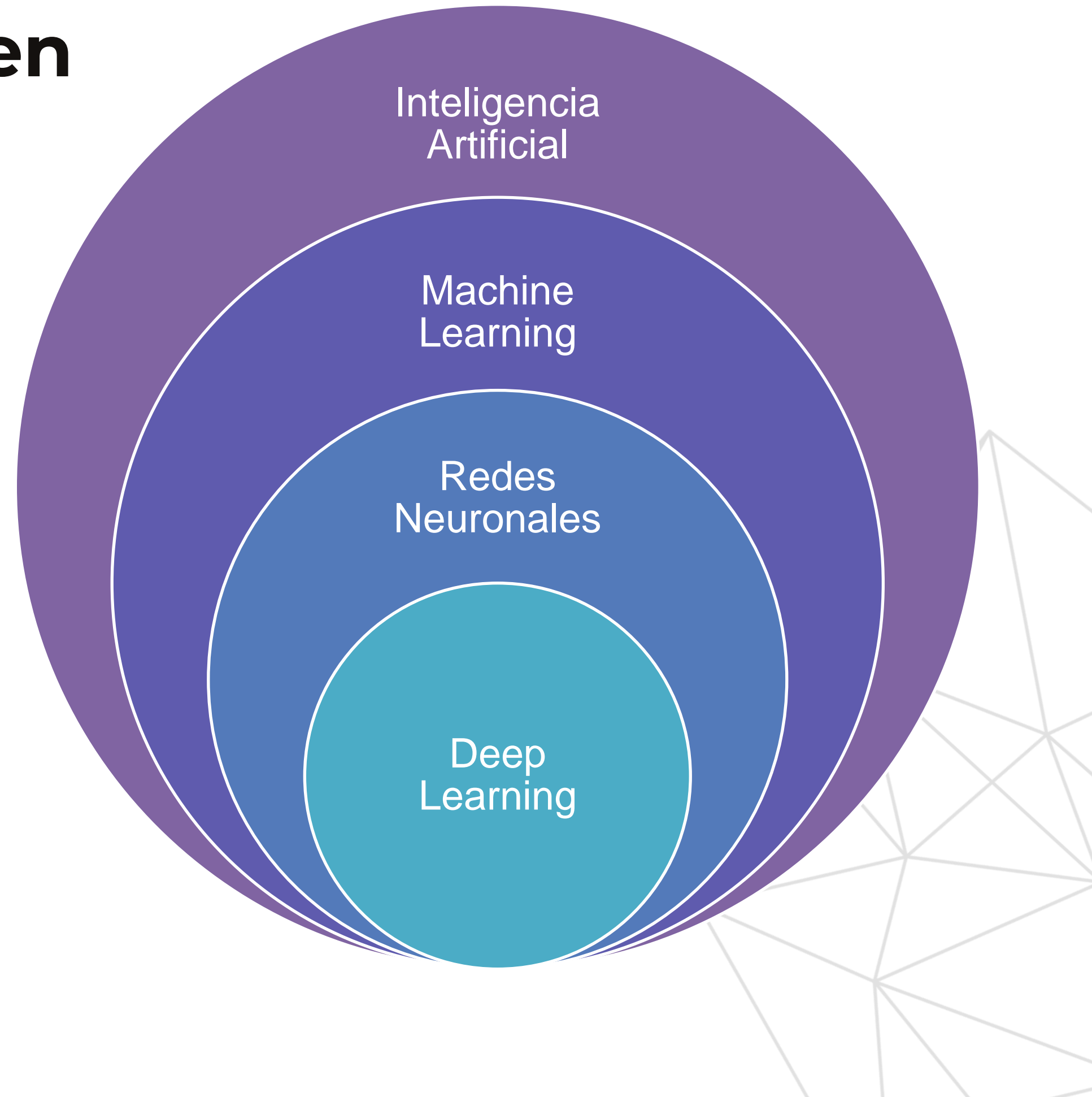
Conceptos clave: Deep Learning

- El aumento de capas en las ANN produce algoritmos cada vez más complejos. Este incremento en el número de capas y en la complejidad es lo que hace que estos algoritmos sean conocidos como técnicas de Deep Learning.



Conceptos clave: Resumen

- Los algoritmos de Deep Learning son un tipo concreto de redes neuronales artificiales (con gran cantidad de capas).
- Las redes neuronales son un tipo de algoritmo dentro de las muchas soluciones existentes en el campo del Machine Learning.
- El Machine Learning es una subdisciplina dentro de la inteligencia artificial que trata de conseguir que las computadores aprendan de la experiencia.

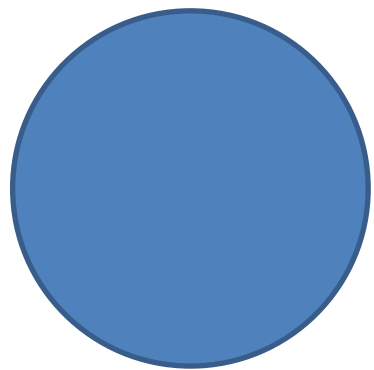
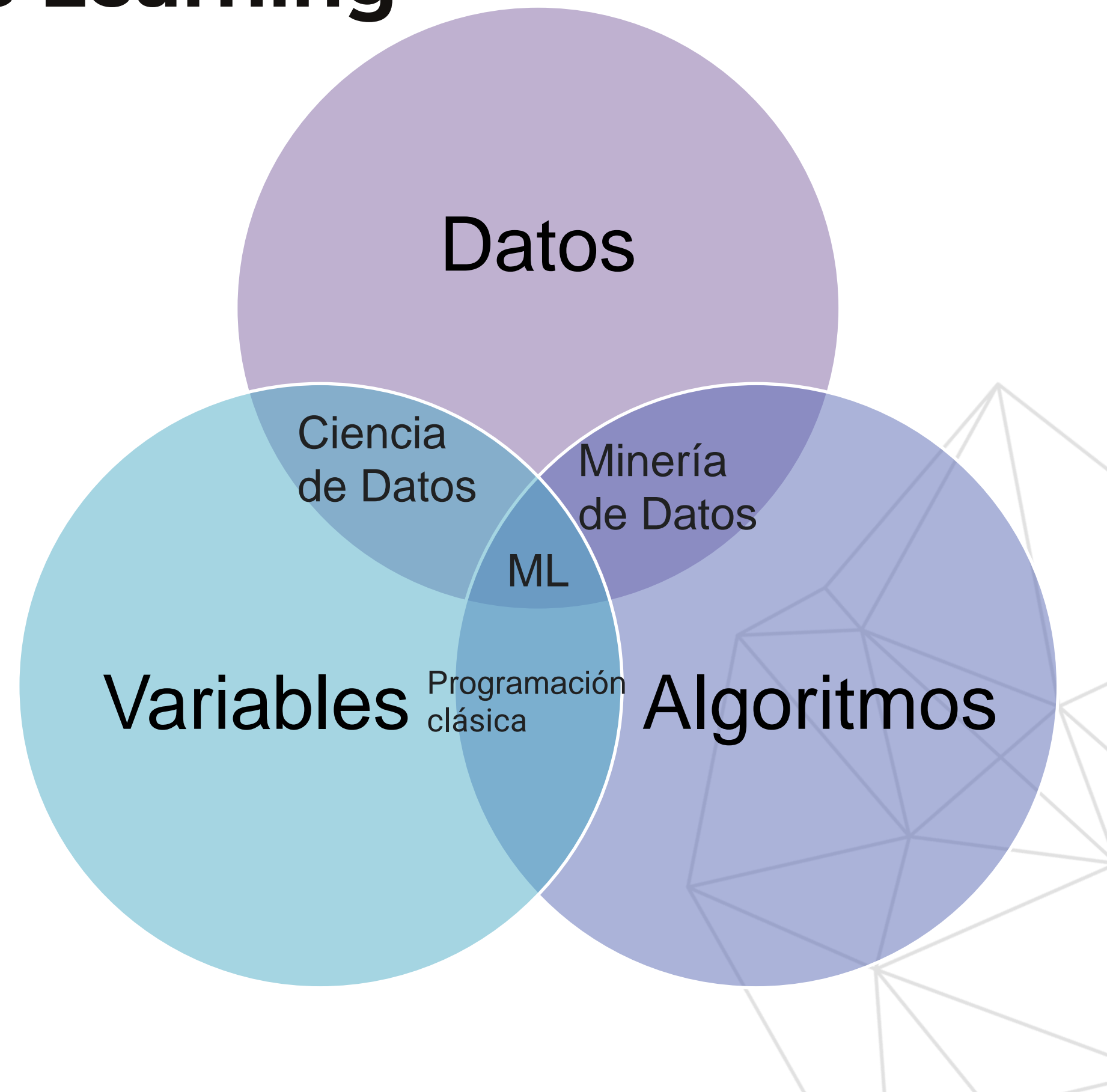


Componentes del Machine Learning

Como hemos visto, el objetivo del Aprendizaje Automático es generar modelos que aprendan, es decir, programas que sean capaces de producir una salida específica dada una entrada concreta.

Por ese motivo, el Machine Learning consta de tres componentes principales:

- Los datos
- Las variables (o *features* en inglés)
- Los algoritmos



Componentes del ML: Datos

Dos formas de recopilar datos:

- Manual
- Automática (también con técnicas de *data augmentation*)

Los datos utilizados para entrenar nuestro modelo deben, al menos, cumplir las siguientes propiedades (para evitar *garbage in garbage out*):

- **Exactitud:** el dato debe ser correcto, ya que de lo contrario se obtendrían resultados erróneos.
- **Representatividad:** los datos deben ser representativos del problema al que nos queremos enfrentar. No debemos darle ni más (focalizarnos en los *outliers*) ni menos representatividad de la que realmente tiene cada casuística.
- **Completitud:** la falta de valores (*missing values*) dificulta el aprendizaje.
- **Abundancia:** para entrenar nuestros modelos necesitamos gran cantidad de datos.

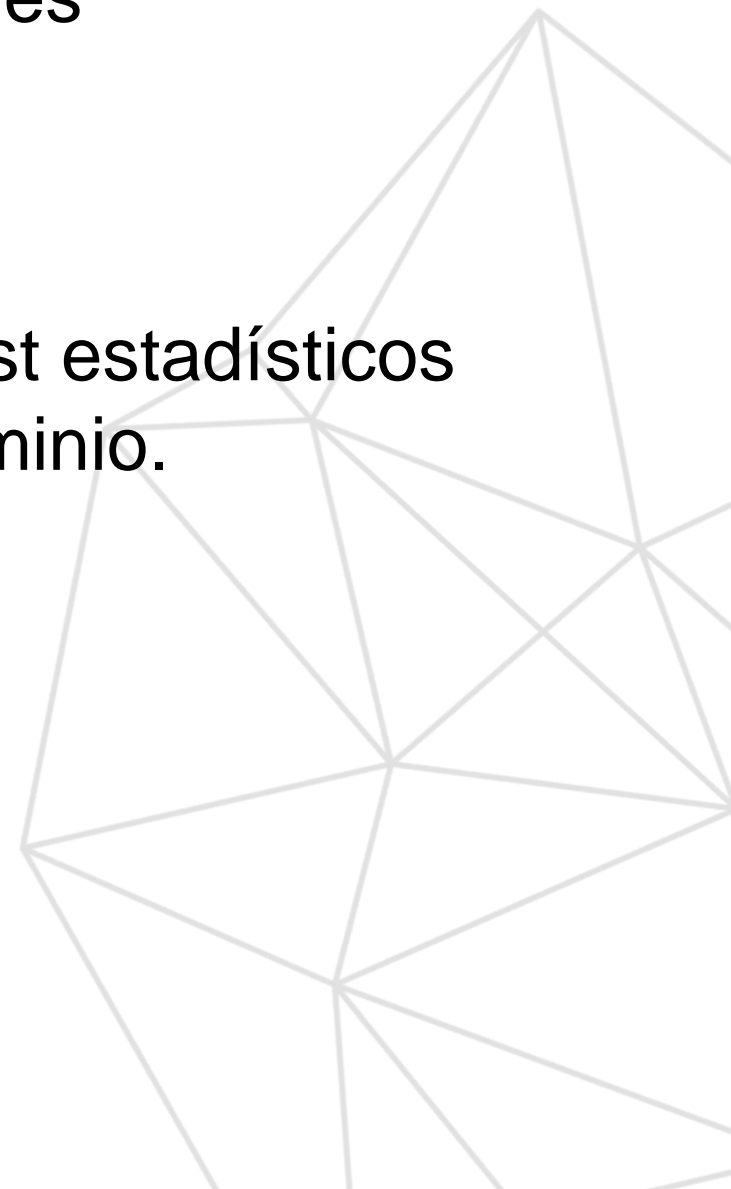


Componentes del ML: Features

Son las variables o parámetros de nuestros datos. Utilizando nuestro conocimiento del dominio y con un análisis exploratorio de los datos (EDA) debemos evitar que las variables elegidas sean:

- **Irrelevantes:** impactarán negativamente en el proceso de aprendizaje.
- **Redundantes:** distintos parámetros deben darnos información distinta (evitar variables correlacionadas).
- **Incompletas:** los parámetros elegidos de nuestros datos deben estar completos.

Para lograrlo podemos usar técnicas como el coeficiente de correlación de Pearson, test estadísticos (como el de la chi cuadrado) u otros métodos además del conocimiento experto del dominio.

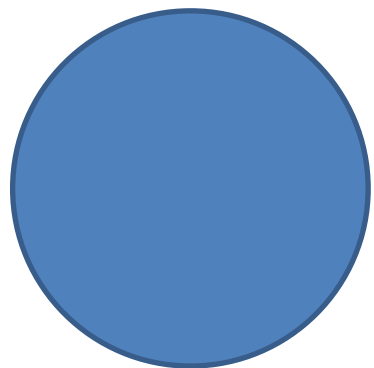


Componentes del ML: Algoritmos

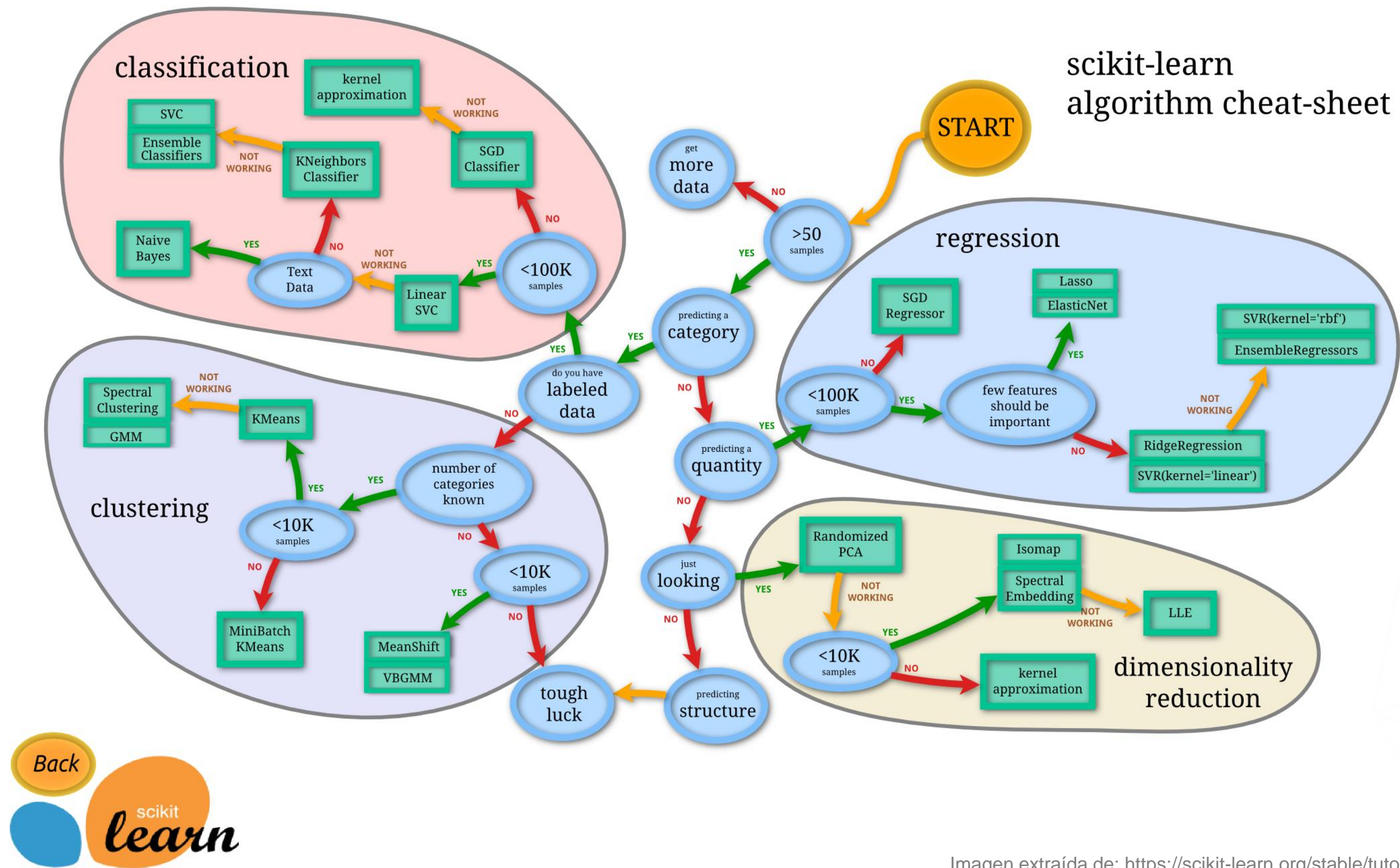
Cada problema es diferente y una de las claves para resolverlo como un buen científico de datos es **elegir el algoritmo que mejor se adecúa a nuestras necesidades**. Esta elección afectará al:

- Rendimiento
- Precisión
- Tamaño
- Eficiencia
- Tiempo de entrenamiento
- Explicabilidad

de nuestro modelo. Es importante conocer el problema, nuestros datos y variables para elegir correctamente el algoritmo necesario.



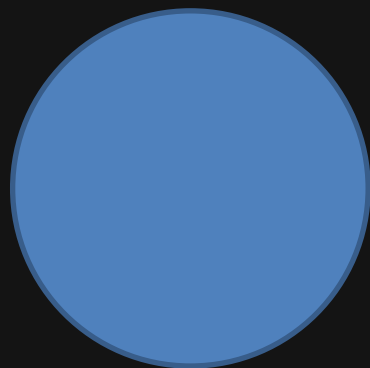
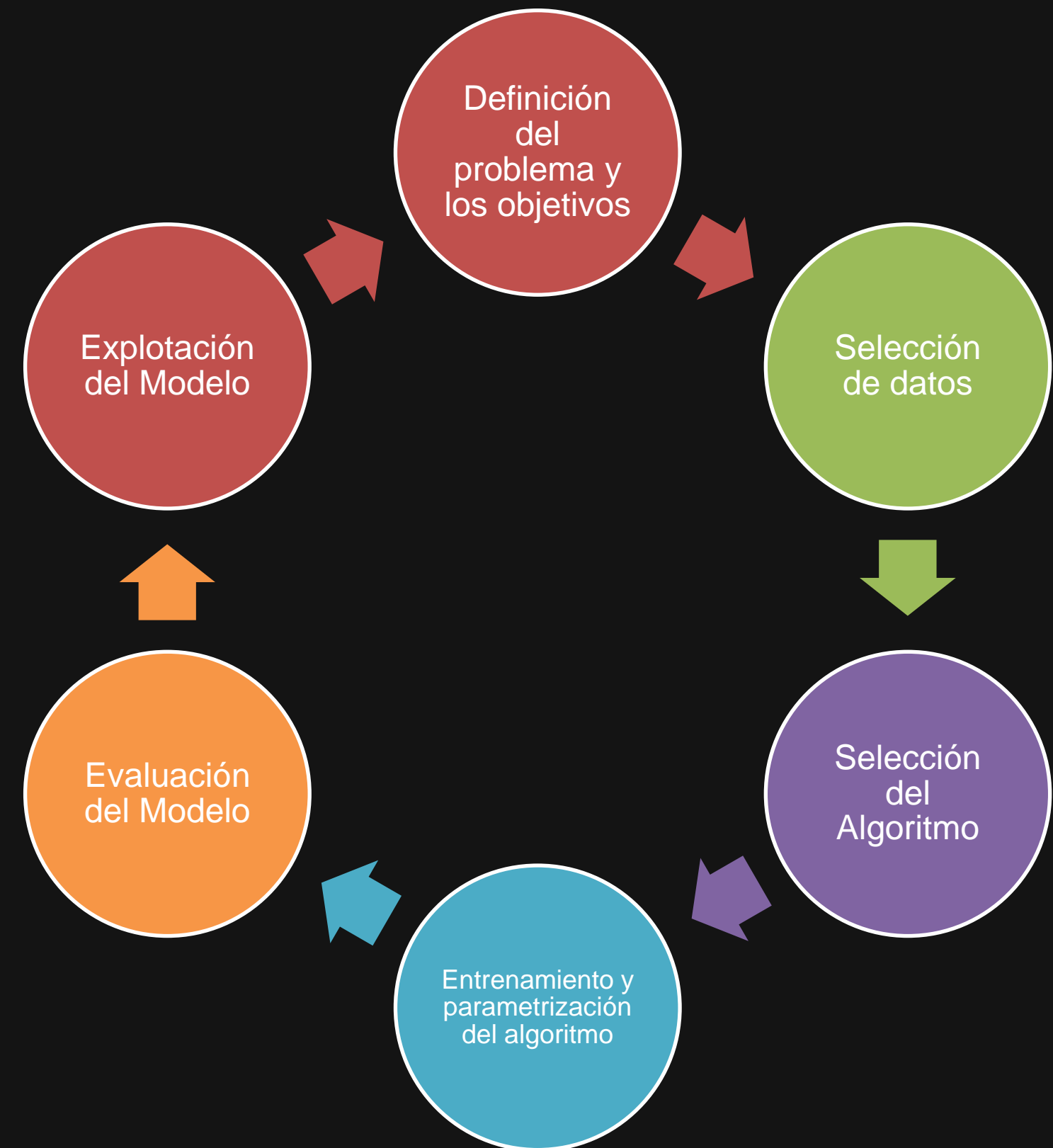
Componentes del ML: Algoritmos



02

Metodología del ML

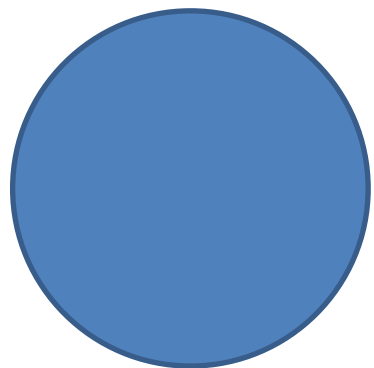
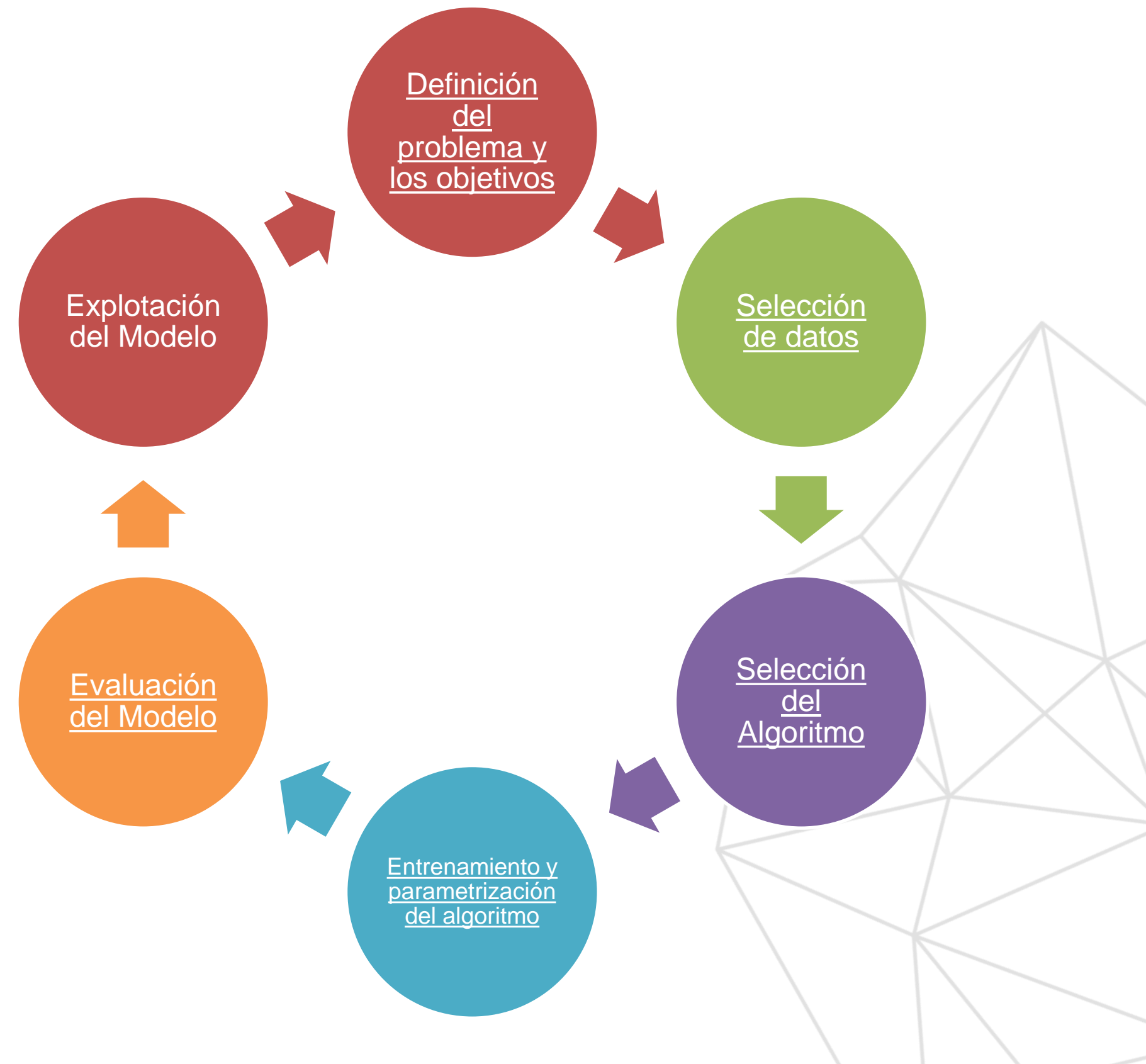
Fases de la metodología para obtener modelos de Aprendizaje Automático



Esquema general del Machine Learning

El propósito de resolver un problema con aprendizaje automático es el de **crear sistemas** que sean **capaces de aprender** por ellos mismo sin ser programados de forma explícita **y aplicar lo aprendido en un entorno real** para la predicción, toma de decisiones, etc.

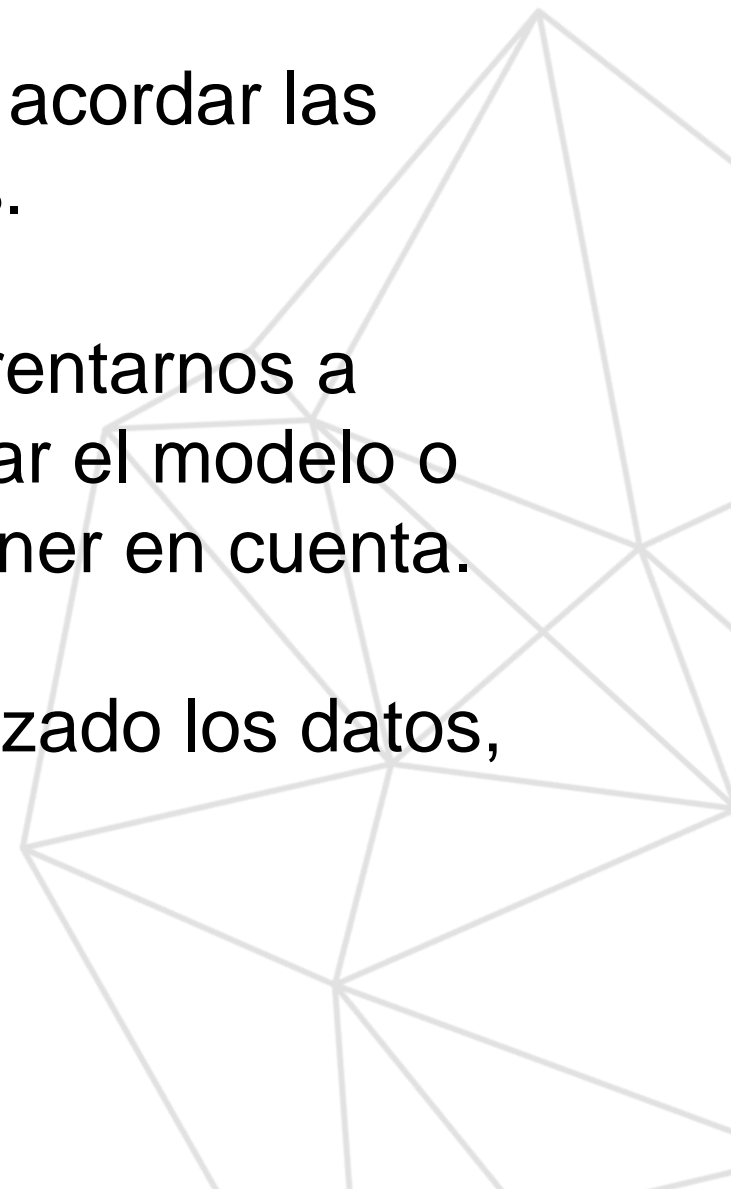
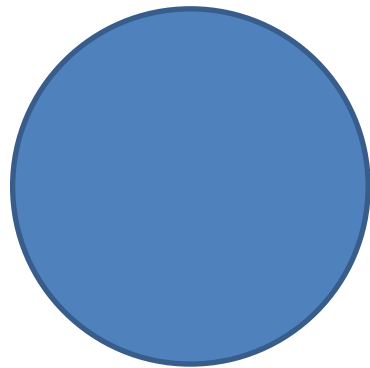
Por tanto podemos decir que el objetivo final es el de **crear el mejor modelo posible para poder explotarlo**. Para lograrlo, podemos seguir la siguiente metodología:



Esquema general del ML: Definición

Una vez conocemos el negocio y los datos disponibles, en la fase de definición del problema y los objetivos debemos resolver las siguientes preguntas:

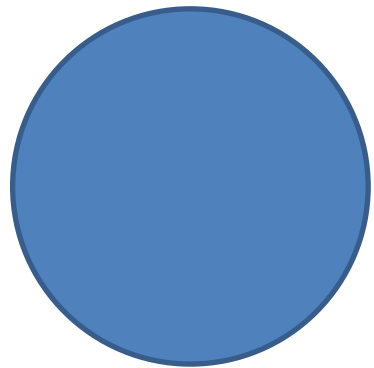
- ¿Qué queremos conseguir con el modelo? Puede ser predecir un valor, clasificar datos en categorías, agrupar conjuntos de datos por similitud, encontrar patrones, ...
- ¿Qué consideraremos como un modelo aceptable? Para ello debemos acordar las métricas que vamos a utilizar y los valores a alcanzar en cada una de ellas.
- ¿Qué restricciones posee nuestro problema? Podríamos tener que enfrentarnos a restricciones de almacenamiento, memoria o velocidad a la hora de ejecutar el modelo o entrenarlo. También existen limitaciones de *explicabilidad* que debemos tener en cuenta.
- ¿Con qué datos contamos para resolver el problema? Tras haber analizado los datos, debemos conocer qué tipo de aprendizaje podemos utilizar.



Esquema general del ML: Tipos de problema

En función del problema al que nos enfrentemos nuestro modelo deberá alcanzar distintos objetivos. El problema puede ser de:

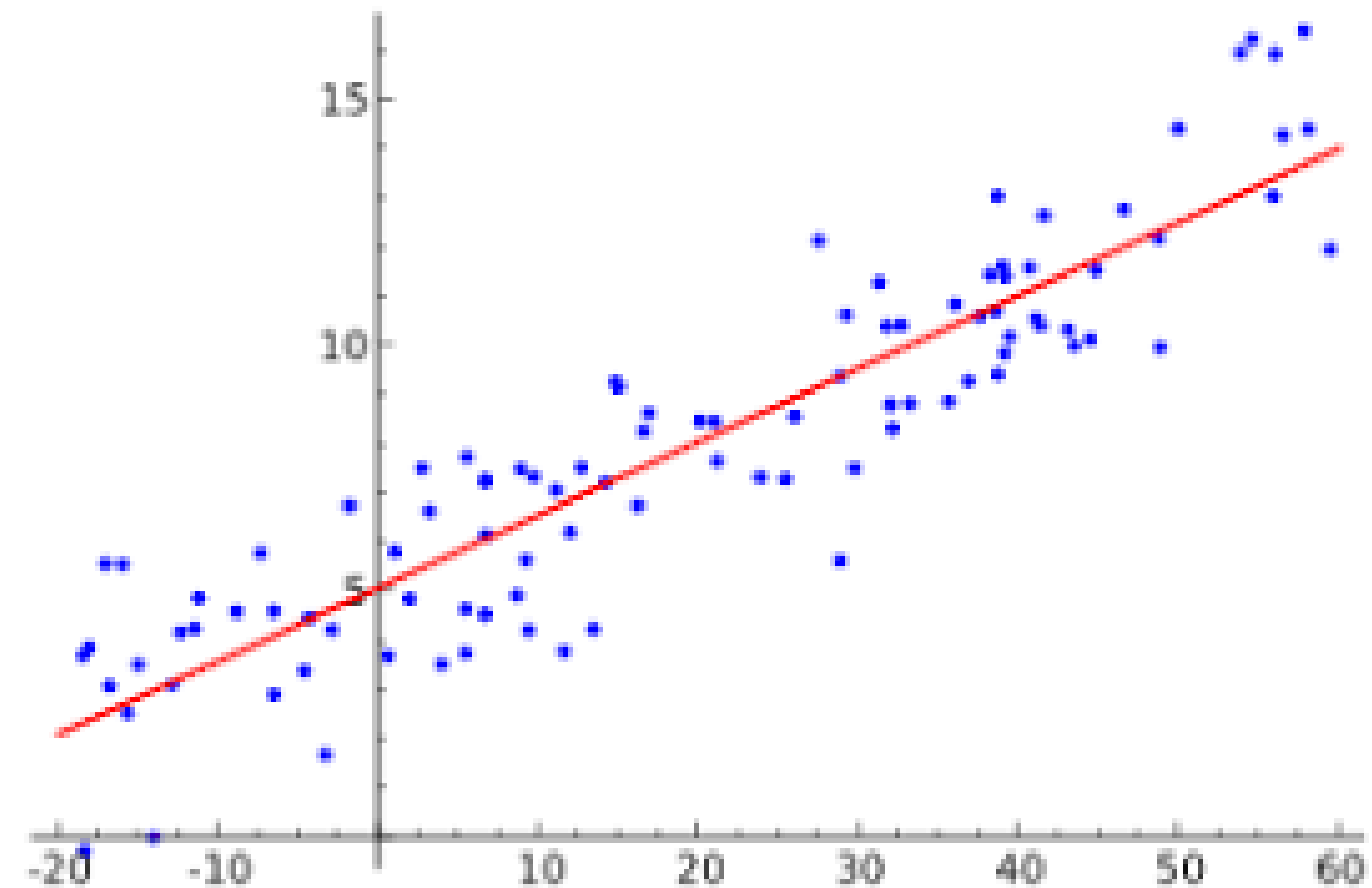
- **Regresión**
- **Clasificación**
- **Clustering**
- **Reconocimiento de patrones**
- **Reducción de la dimensionalidad**



Tipos de problema de ML: Regresión

El objetivo final de un problema de regresión es el de **predecir un valor continuo**. Este tipo de problemas se caracterizan por:

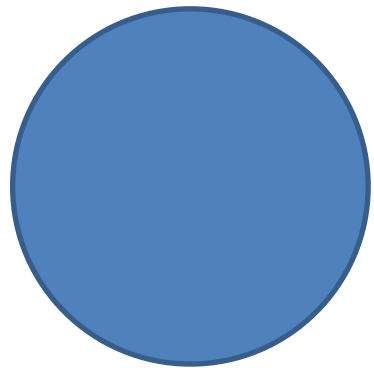
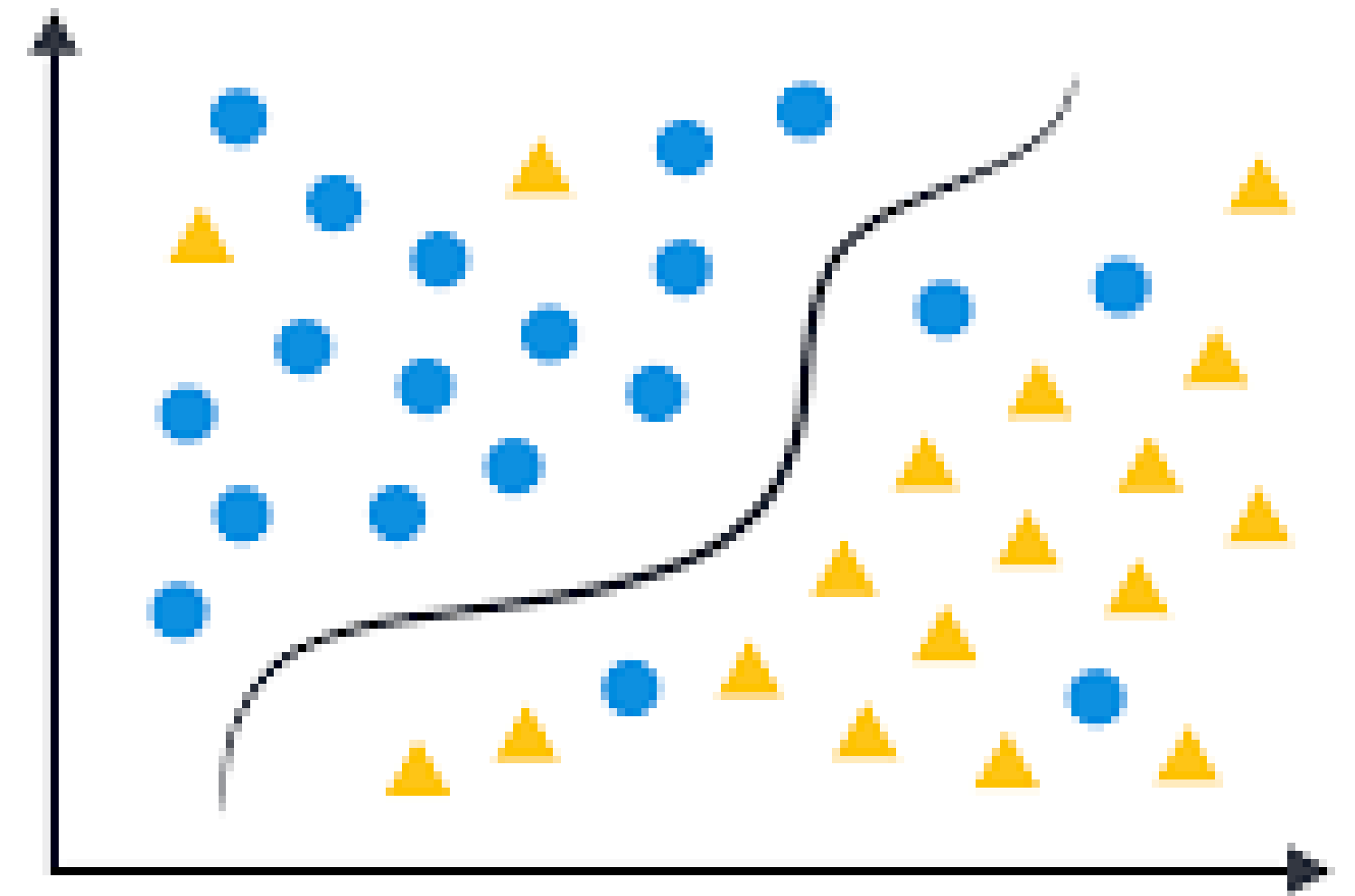
- Datos etiquetados previamente
- Salida de valores continuos
- Minimización de una función de error entre el valor predicho y el real.
- Ejemplos: predicción de precios, de parámetros biológicos, evolución de comportamiento, ...



Tipos de problema de ML: Clasificación

El objetivo final de un problema de clasificación es el de **clasificar cada elemento en alguna de las categorías existentes**. Este tipo de problemas se caracterizan por:

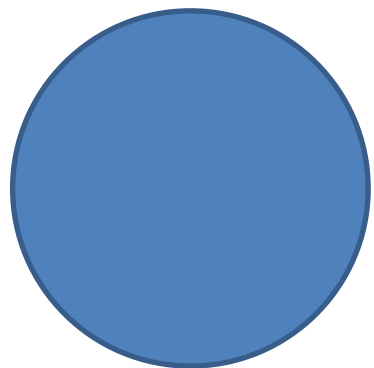
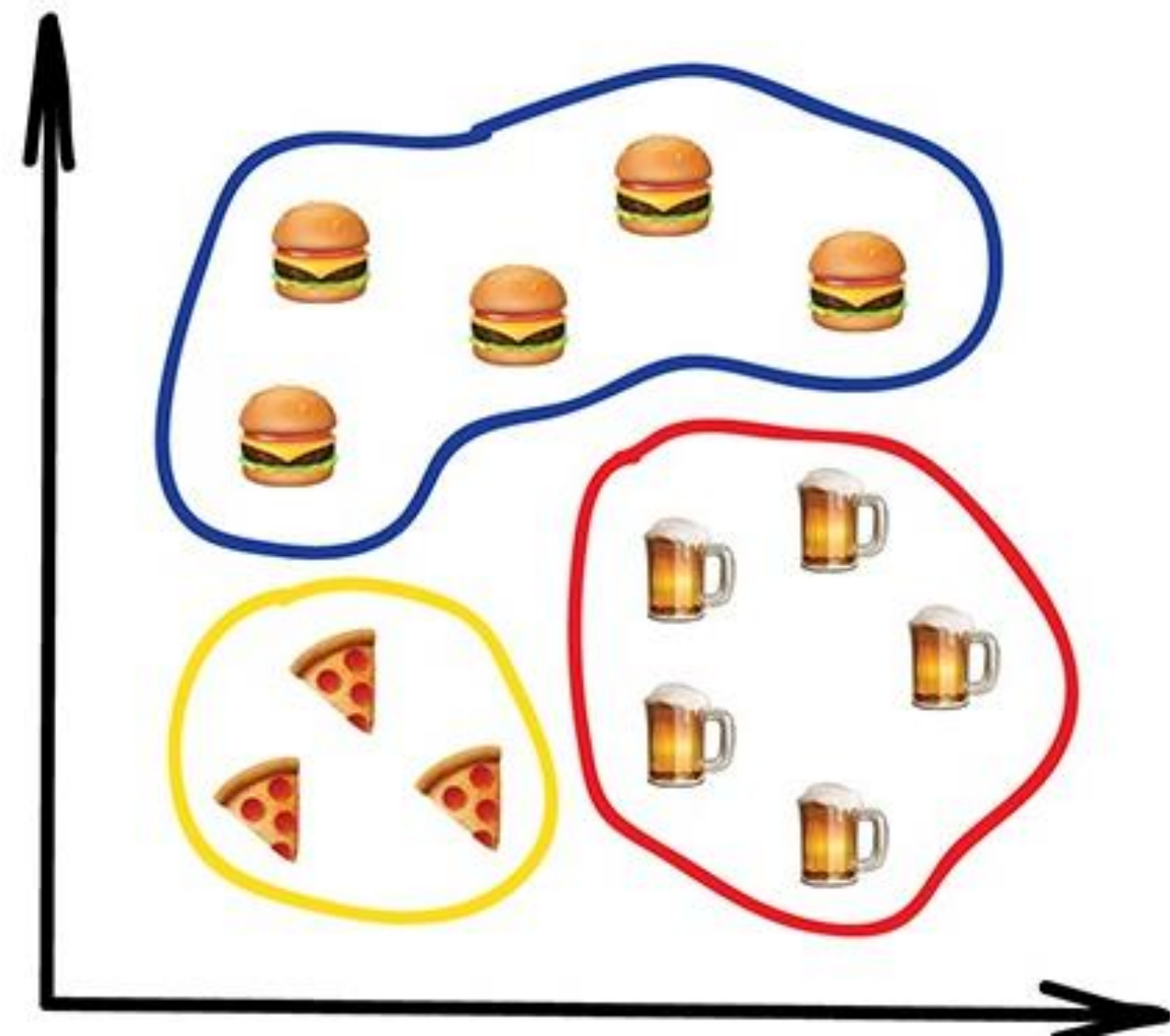
- Datos etiquetados previamente
- Salida de valores discretos preestablecidos
- Minimización de una función de error entre la clase predicha y la real.
- Ejemplos: detección de spam, del idioma, de fraude, concesión de préstamos, análisis de sentimientos, clasificación de imágenes, ...



Tipos de problema de ML: Clustering

El objetivo final de un problema de clustering es el de **agrupar por similitud elementos no etiquetados**. Este tipo de problemas se caracterizan por:

- Datos no etiquetados previamente
- Salida de valores discretos
- Minimización de métricas propias de los clusters que miden su calidad y consistencia.
- Ejemplos: segmentación de mercado, compresión de imágenes, detección de anomalías, ...

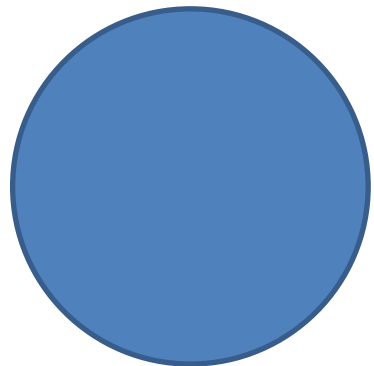
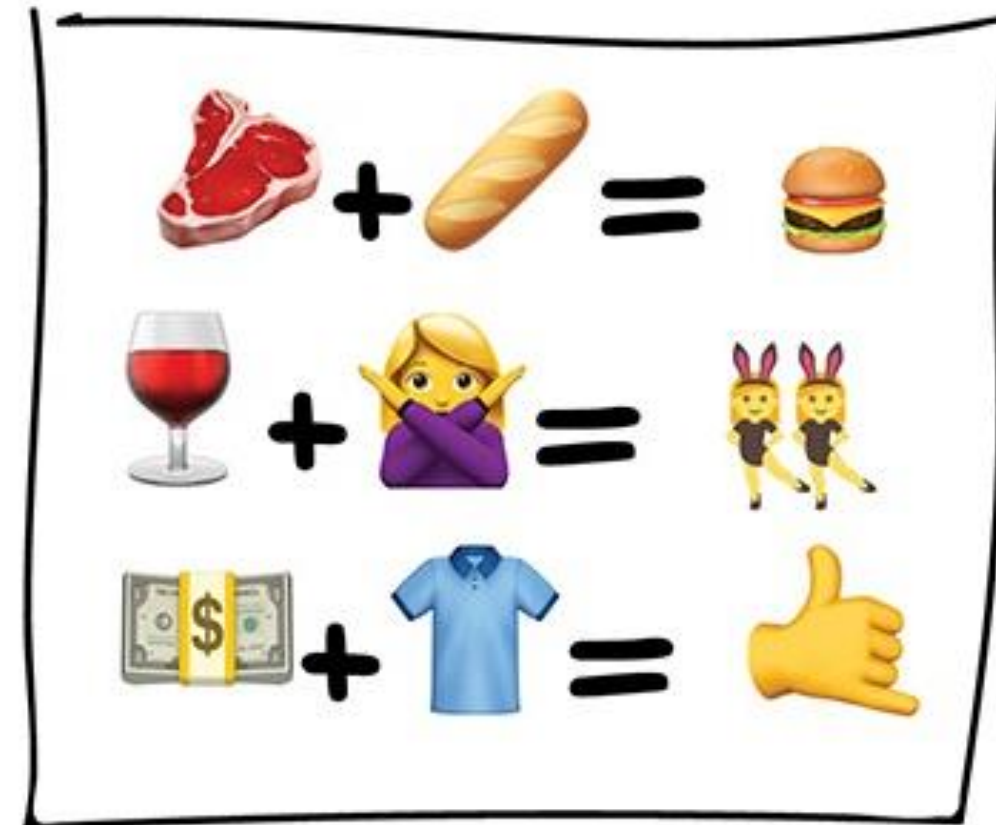


Tipos de problema de ML: Patrones

El objetivo final de un problema de reconocimiento de patrones (o reglas de asociación) es el de **extraer propiedades de subconjuntos del conjunto de datos**. Este tipo de problemas se caracterizan por:

- Relaciones no generadas previamente
- Salida de valores discretos
- Maximización de una función de confianza de las reglas obtenidas.
- Ejemplos: predicción de ventas, análisis de patrones de navegación web, ...

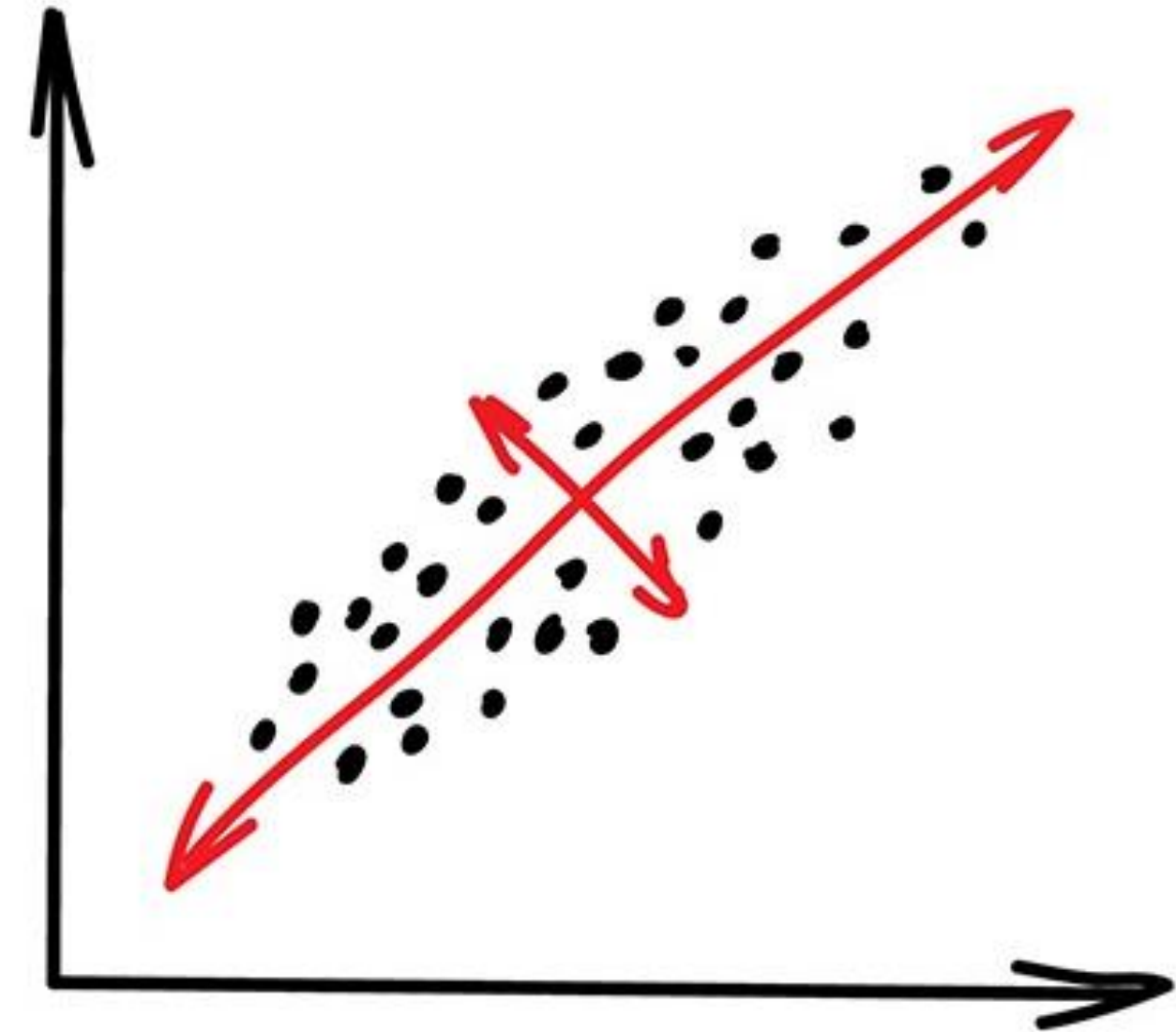
$$\{cebollas, vegetales\} \Rightarrow \{carne\}$$



Tipos de problema de ML: Dimensionalidad

El objetivo final de un problema de reducción de dimensionalidad es el de **reducir el número de variables utilizadas y simplificar el problema**. Este tipo de problemas se caracterizan por:

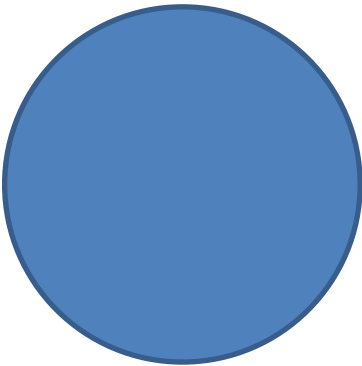
- Salida de valores discretos
- Minimización de métricas de relación entre las distintas variables
- Ejemplos: clasificación de documentos por temas, búsquedas semánticas, ...



Tipos de problema de ML: Resumen

Para contestar a la pregunta “¿Qué queremos conseguir con el modelo?” en la fase de definición deberemos identificar el tipo de problema al que nos enfrentamos:

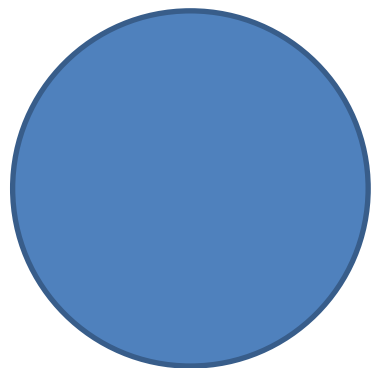
Tipo de problema	Información previa de la salida	Tipo de salida	Objetivo
Regresión	Sí	Continua	Minimizar error
Clasificación	Sí	Discreta	Minimizar error
Clustering	No	Discreta	Maximizar calidad y consistencia
Reconocimiento de patrones	No	Discreta	Maximizar confianza
Reducción de dimensionalidad	No	Discreta	Minimizar relación



Esquema general del ML: Aceptación del modelo

Tras responder a la pregunta “¿Qué queremos conseguir con el modelo?” debemos contestar a “¿Qué consideramos como un modelo aceptable?”. Para ello debemos elegir las métricas de evaluación del modelo que utilizaremos (y explicaremos más adelante) y determinar los valores mínimos (o máximos) a superar (o a reducir). Para seleccionar estos valores podemos partir de:

- Acuerdo experto entre el usuario y el equipo de científicos de datos.
- Rendimiento o eficacia de un modelo anterior.
- Usar modelos *Dummy* (en la librería *scikit-learn* encontramos [implementaciones](#) de estos estimadores para problemas de regresión y clasificación)



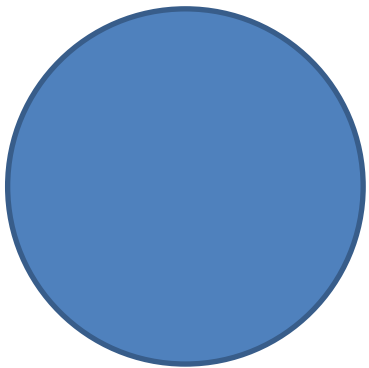
Esquema general del ML: Tipos de aprendizaje

Otra de las preguntas que debemos hacernos en la fase de definición del problema y los objetivos, es “¿Con qué datos contamos para resolver el problema?”. Y es que, esta información es crucial para un proyecto de Ciencia de Datos.

El tipo de información de la que disponemos, como se ha visto, **influye en el tipo de problema** ante el que nos encontramos.

Por otro lado, **influye en el tipo de aprendizaje que podemos utilizar**. Existen los siguientes tipos:

- [Aprendizaje Supervisado](#)
- [Aprendizaje No Supervisado](#)
- [Aprendizaje por Refuerzo](#)

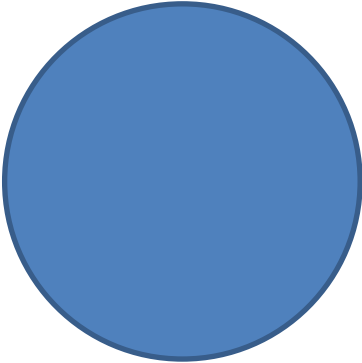


Tipos de aprendizaje: Supervisado

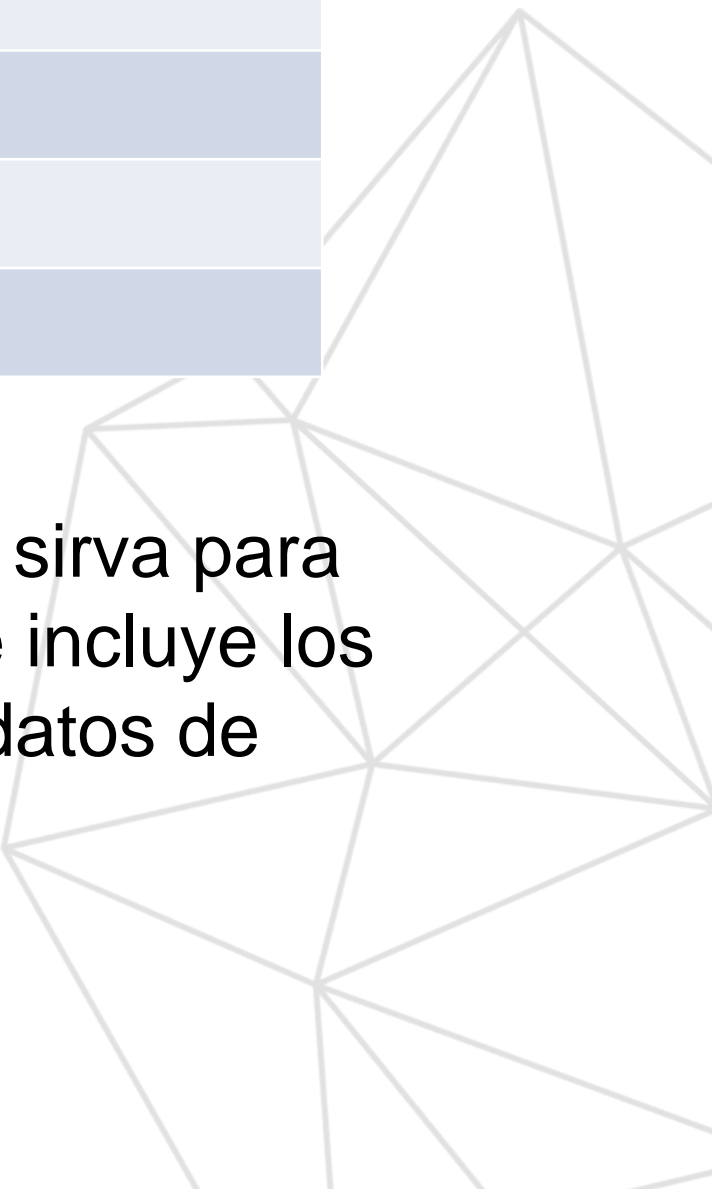
Es el tipo de aprendizaje que se basa en **descubrir la relación** existente **entre** unas variables de **entrada** y unas variables de **salida**. Por lo tanto, para este aprendizaje se caracteriza por:

- Requiere datos etiquetados para el entrenamiento
- Mejores resultados debido a mayor información
- Resuelve problemas de regresión y clasificación

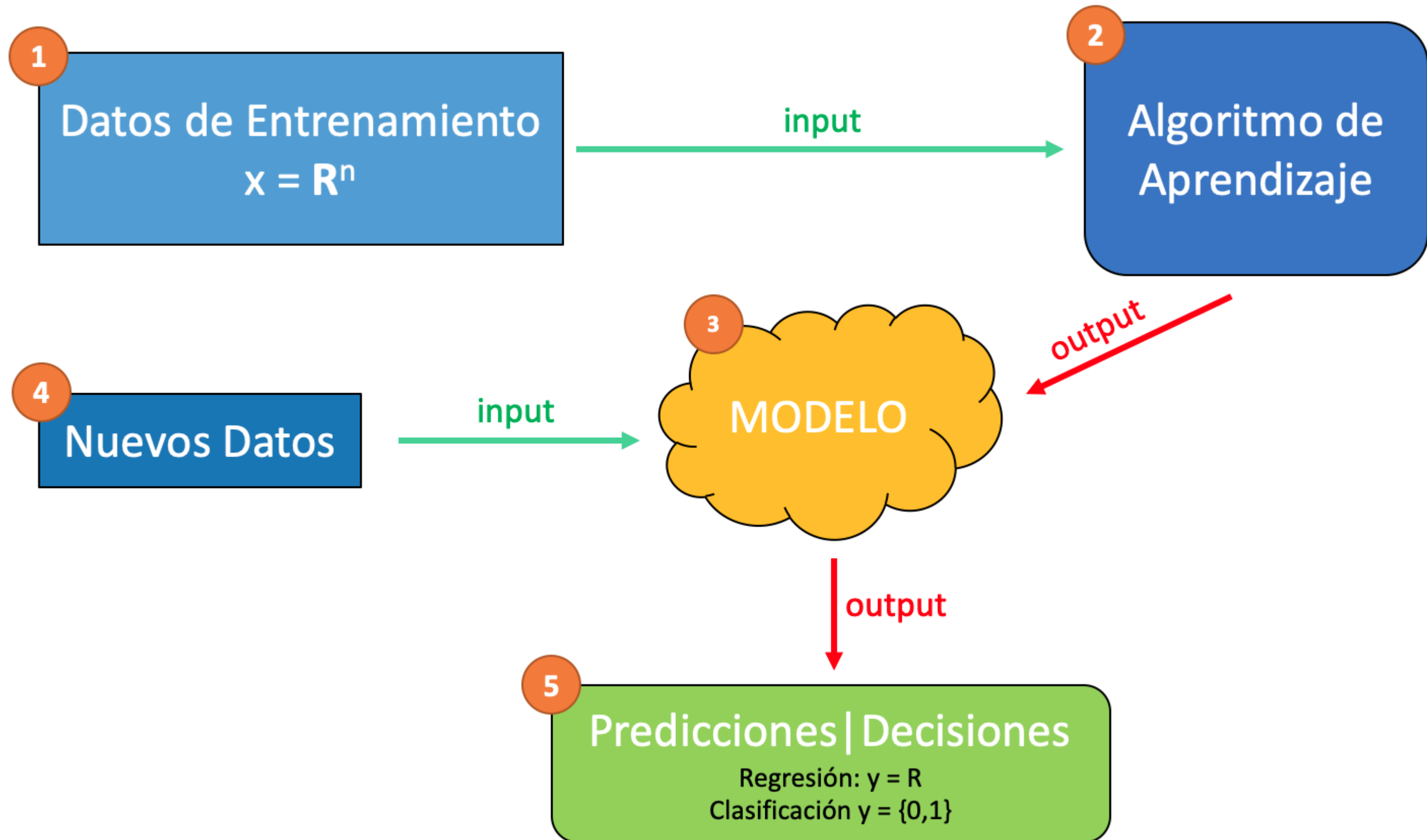
Entrada	Salida
1	2
2	4
6	12
9	18
15	?



El algoritmo de aprendizaje será el encargado de generar un modelo que sirva para predecir. Este modelo lo genera minimizando la **Función de pérdida** que incluye los errores cometidos en la predicción de cada observación del conjunto de datos de entrenamiento.



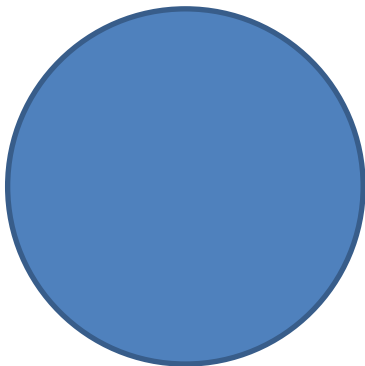
Tipos de aprendizaje: Supervisado















Tipos de aprendizaje: No Supervisado







El aprendizaje no supervisado es el paradigma que consigue **producir conocimiento únicamente de los datos proporcionados como entrada**, sin necesidad de explicarle al sistema el resultado que deseamos obtener. Por lo tanto, para este aprendizaje se caracteriza por:

- No requiere datos etiquetados para el entrenamiento
- Resuelve problemas de clustering, reconocimiento de patrones y reducción de la dimensionalidad









Poligonales

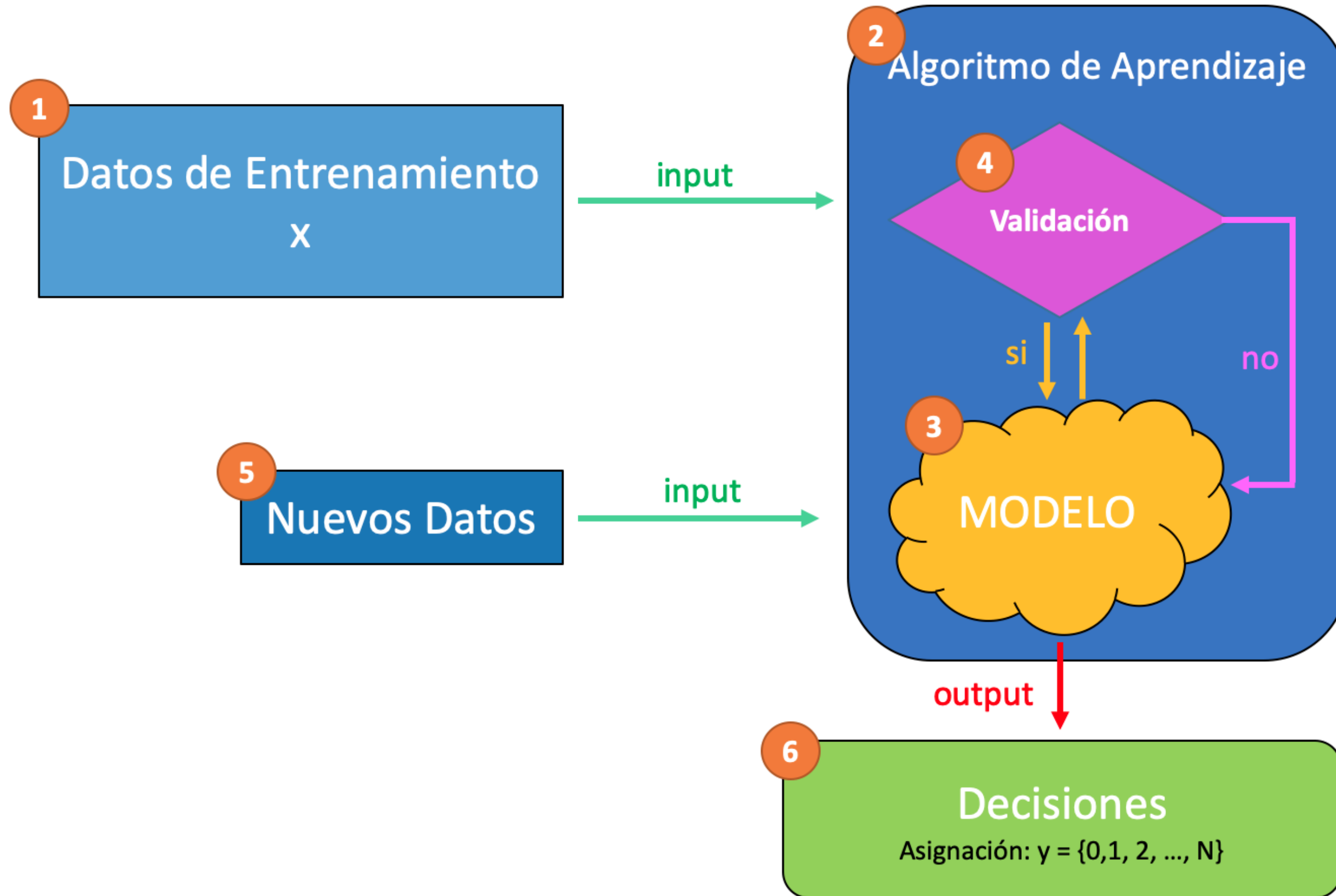
	
	
	

Lineales



Tipos de aprendizaje: No Supervisado

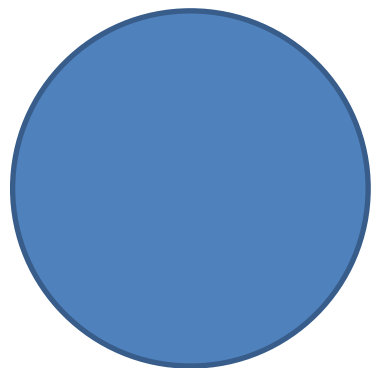
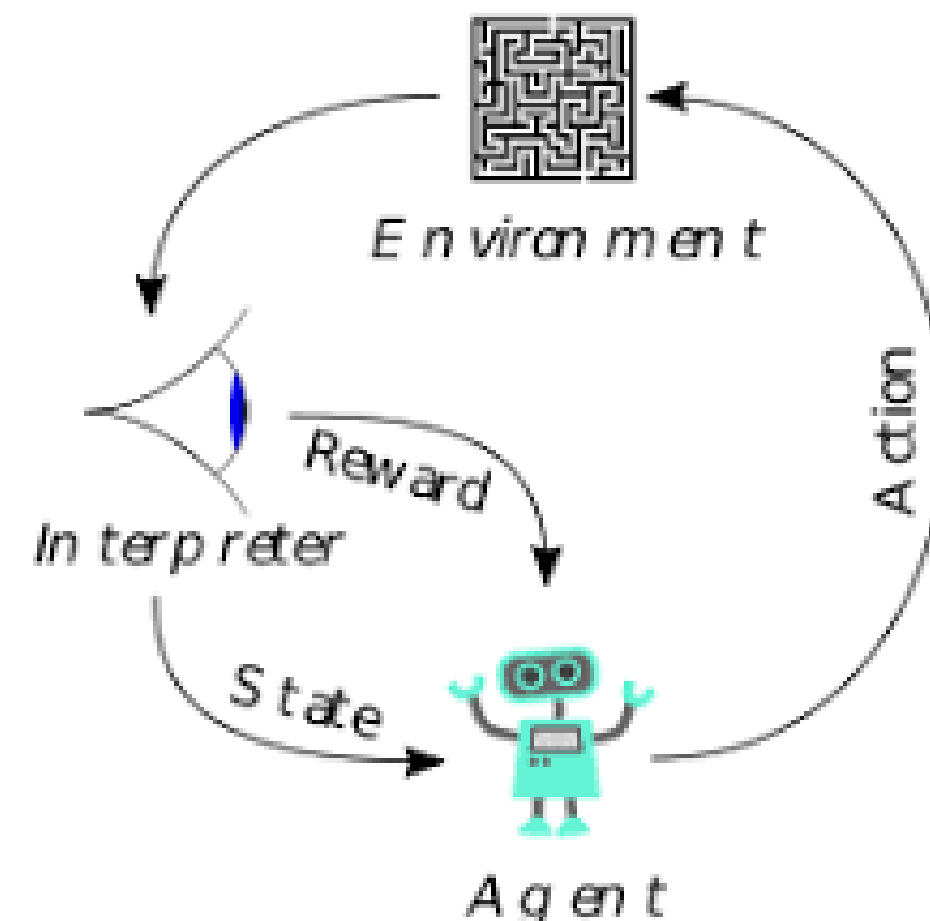


Tipos de aprendizaje: Reforzado

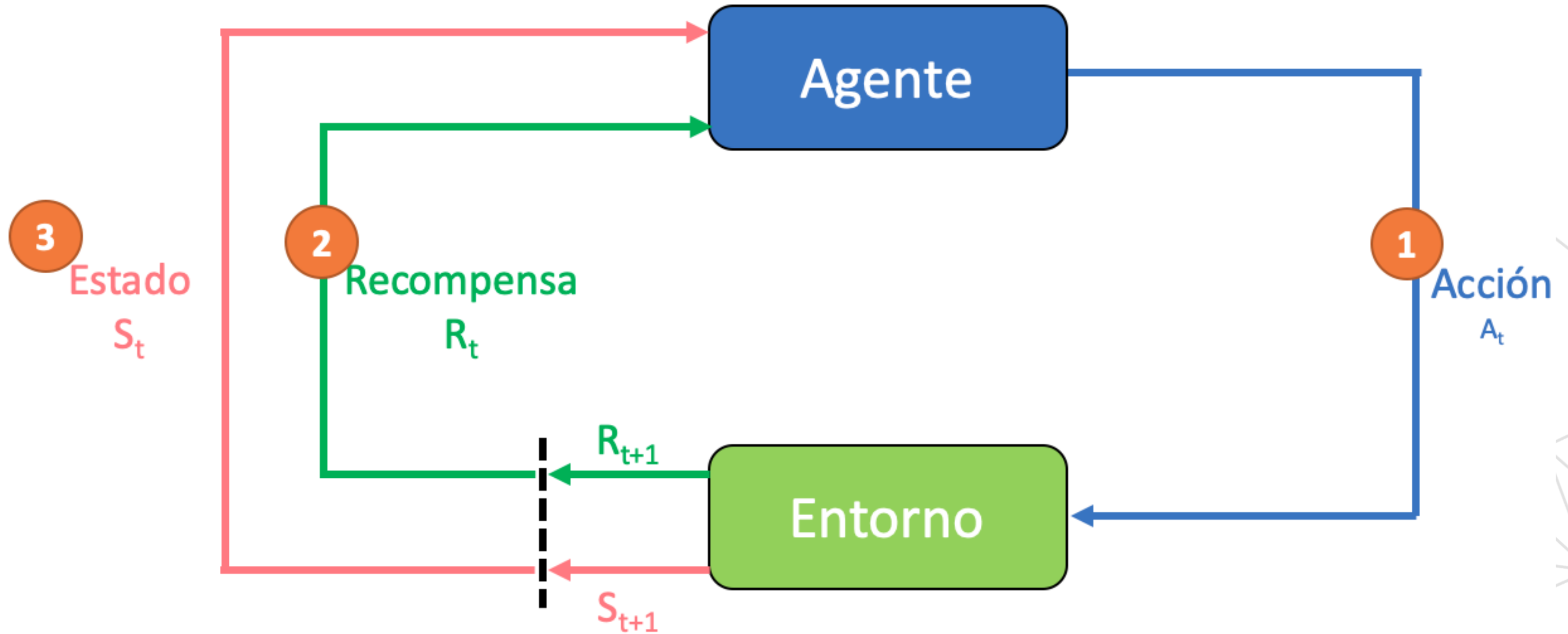
El aprendizaje por refuerzo está basado en un **agente** que "vive" en un **entorno** y es capaz de percibir un **estado** de ese entorno. El agente puede ejecutar **acciones** en cada estado y estas acciones conllevan diferentes **recompensas**.

El objetivo es aprender una **política** encargada de tomar la decisión de qué acción tomar en cada estado con el **objetivo de maximizar la recompensa**, bien sea a corto o largo plazo.

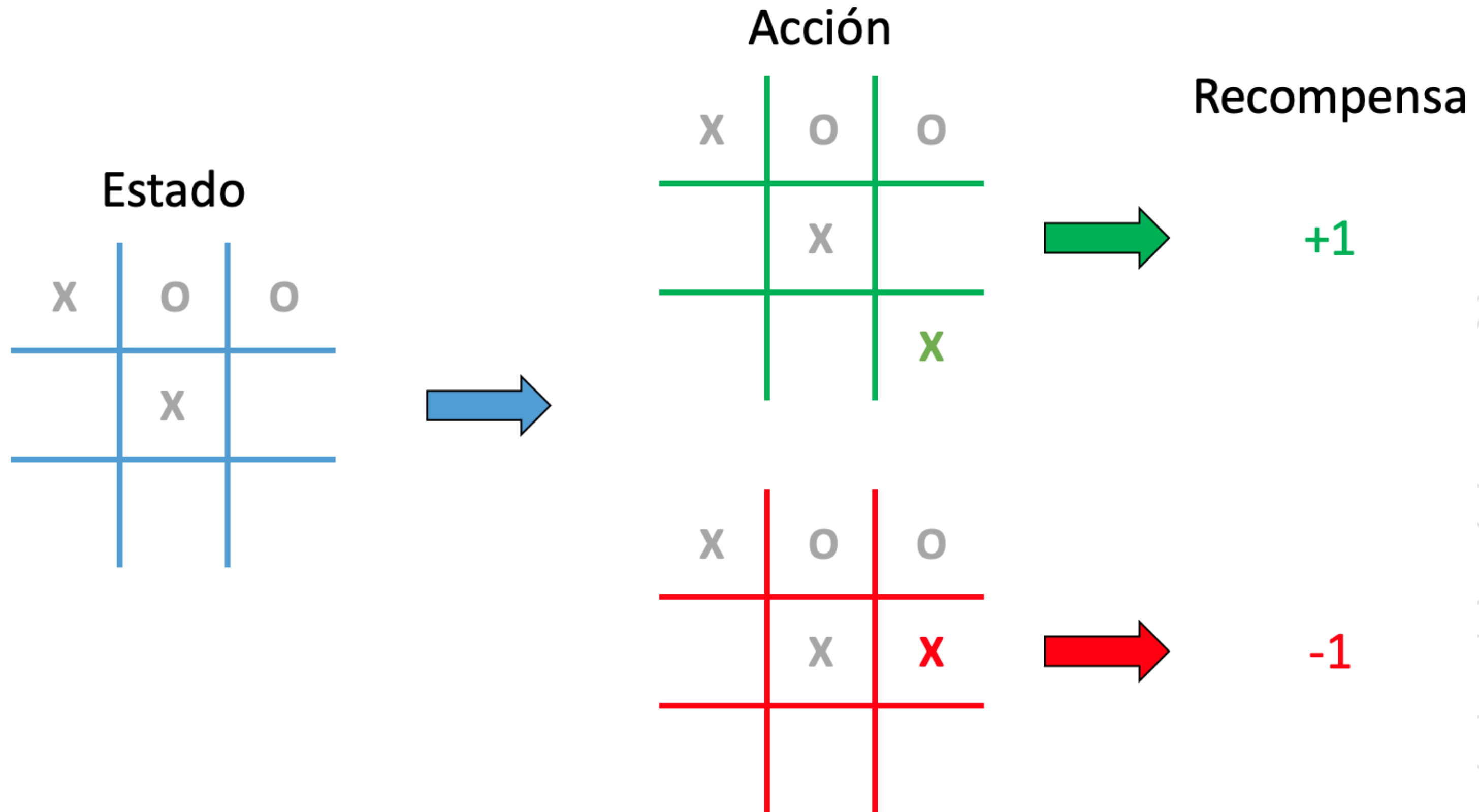
- No requiere datos de ningún tipo para el entrenamiento.
- Necesita un entorno con un sistema de recompensas.
- Buenos resultados para problemas como los juegos.



Tipos de aprendizaje: Reforzado



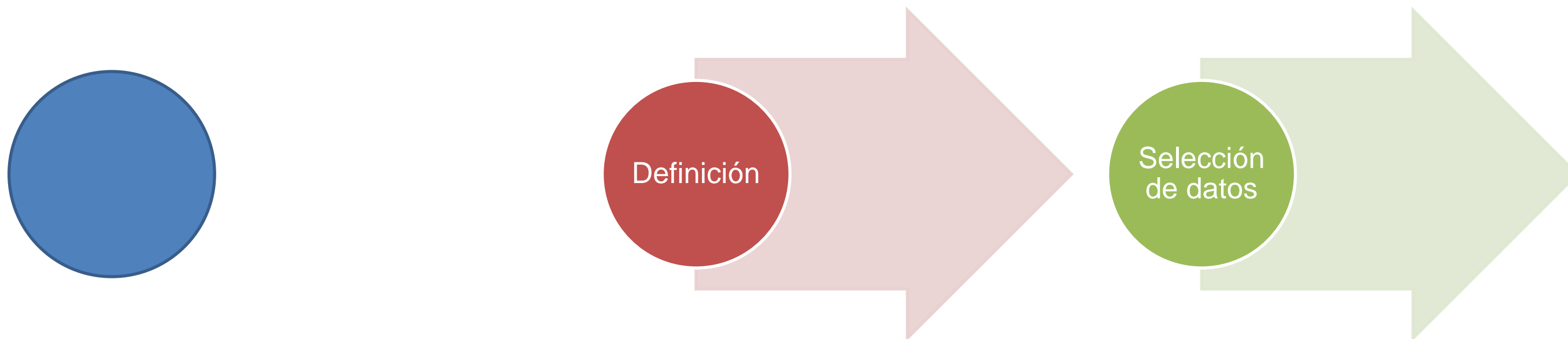
Tipos de aprendizaje: Reforzado



Esquema general del ML: Resumen Definición

En la fase de definición del problema y los objetivos debemos precisar:

- **El tipo de problema de ML:** regresión, clasificación, clustering, reconocimiento de patrones o reducción de la dimensionalidad.
- **Aceptación del modelo:** las métricas y los valores para considerar un modelo aceptable.
- **Restricciones y limitaciones de nuestro problema:** almacenamiento, memoria, velocidad, explicabilidad, ...
- **Tipo de aprendizaje:** supervisado, no supervisado o por refuerzo.



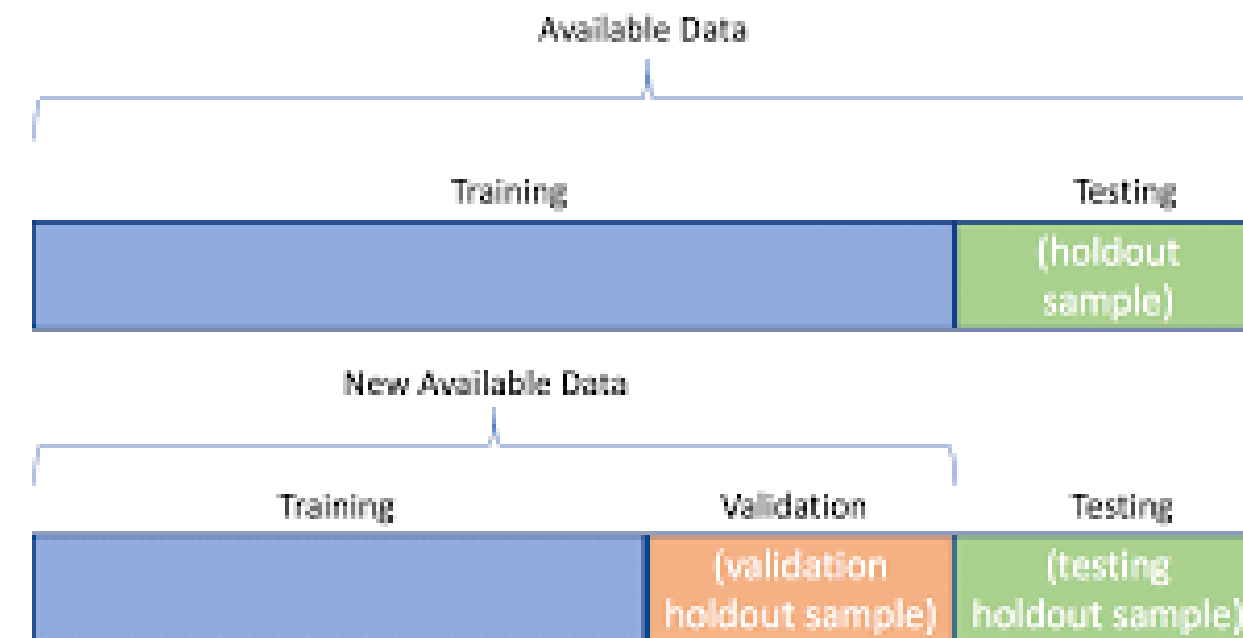
Esquema general del ML: Selección de datos

Tras la fase de definición pasaremos a la fase de Selección de datos. Como hemos visto, para afrontar un problema con Machine Learning es necesario contar con un **dataset suficientemente grande**. Este conjunto de datos, lo tenemos que dividir en:

- **Conjunto de entrenamiento** (*train* en inglés): datos utilizados para entrenar el modelo.
- **Conjunto de datos de test**: datos utilizados para evaluar el rendimiento del modelo.

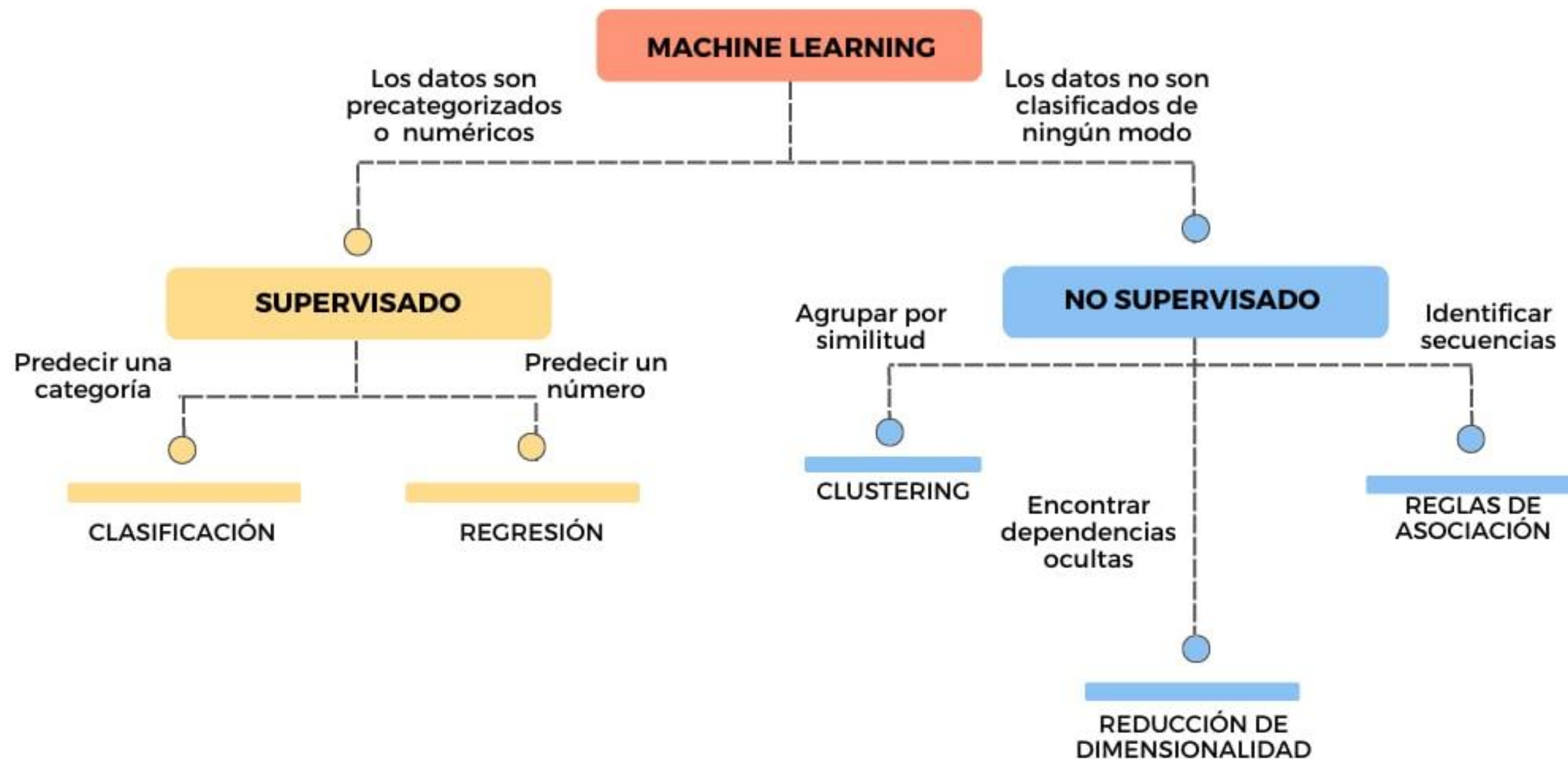
Es importante remarcar que en este paso **no se realizan limpieza de datos** (cuyas técnicas deben aplicarse antes o después según el tipo de limpieza). En ocasiones podemos dividirlo en una tercera parte:

- **Conjunto de datos de validación**: datos utilizados para elegir los hiperparámetros del modelo.

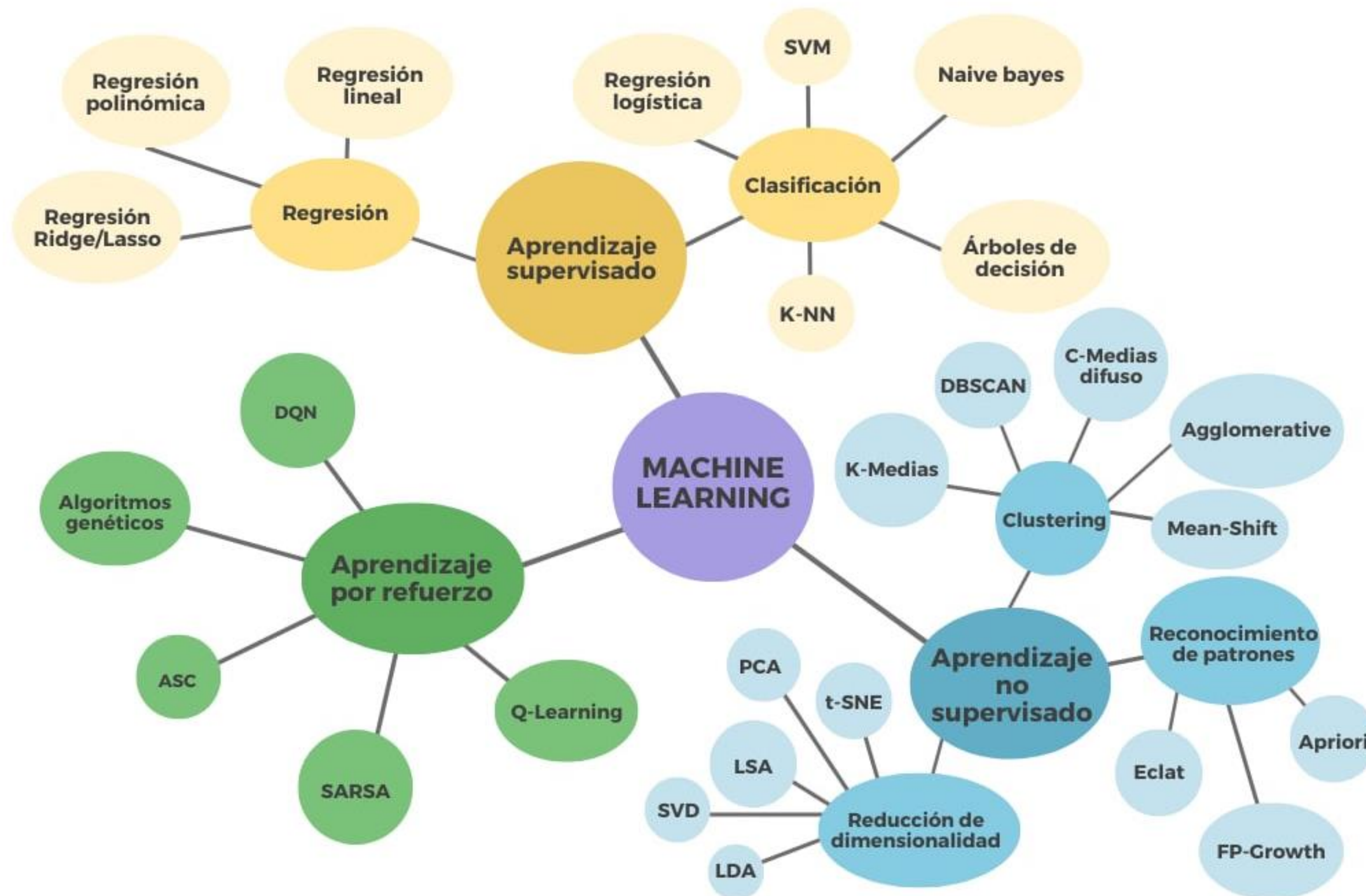


Esquema general del ML: Selección del algoritmo

Tras la fase de Selección de datos pasaremos a la fase de Selección del algoritmo, en función del tipo de problema y tipo de aprendizaje determinados en la fase de definición:



Esquema general del ML: Selección del algoritmo

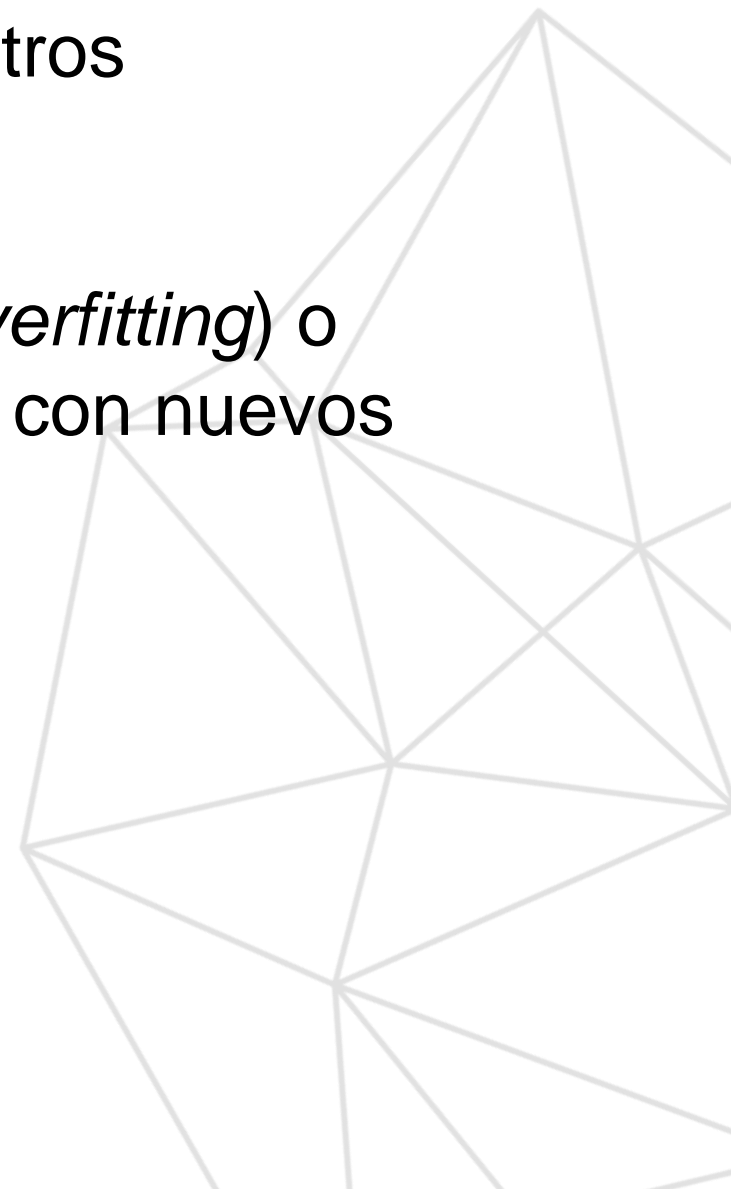
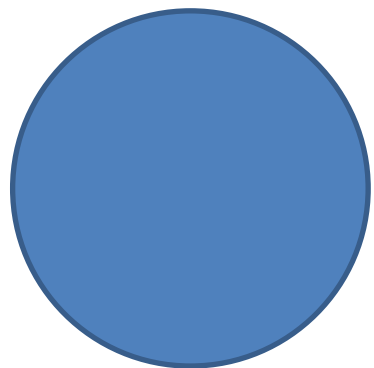


Esquema general del ML: Entrenamiento

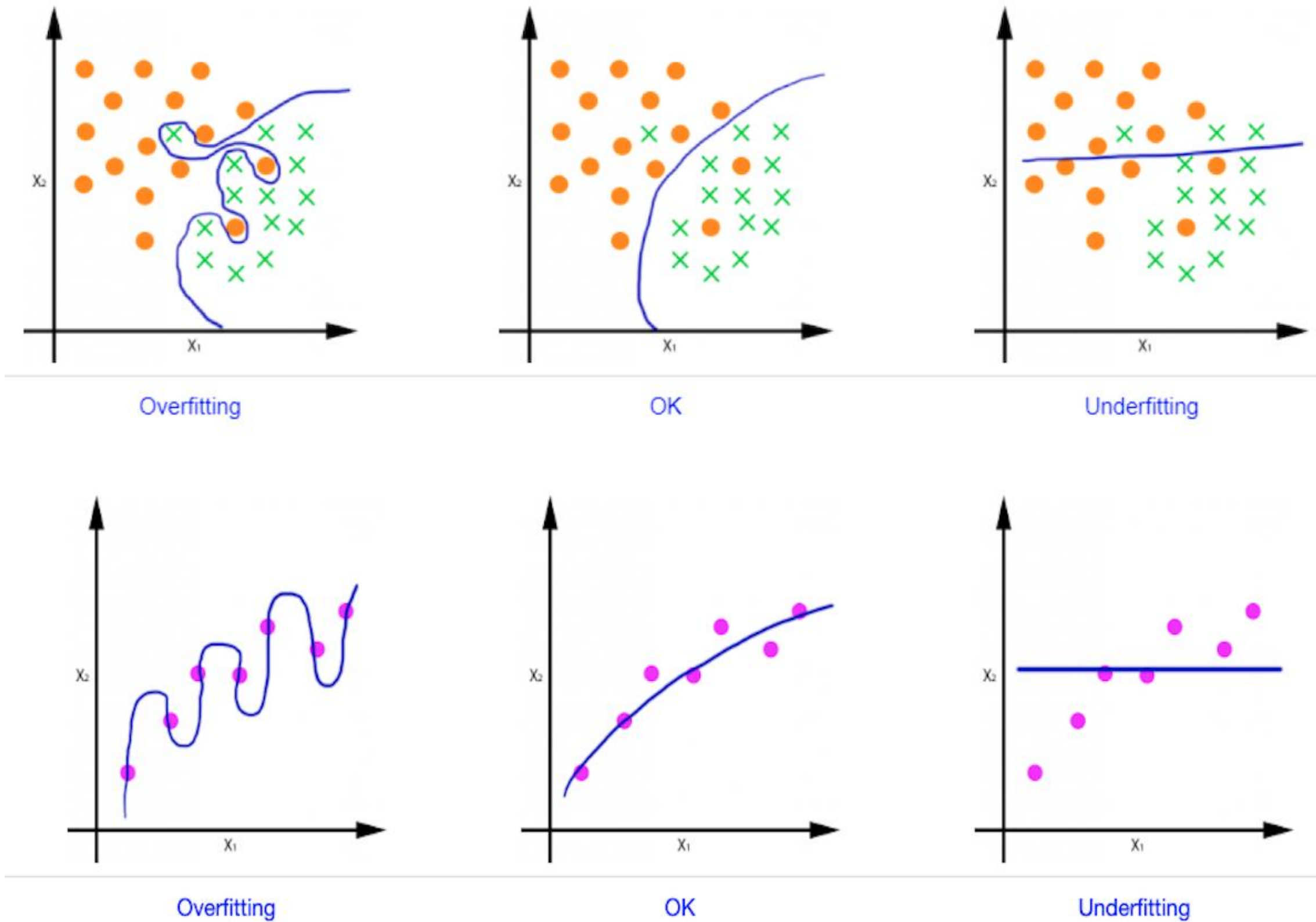
Tras la fase de Selección del algoritmo pasaremos a la fase de Entrenamiento y parametrización del modelo. En caso de que se haya extraído un conjunto de datos para validación se seguirán los siguientes pasos:

1. **Entrenamiento** del modelo con unos hiperparámetros
2. Cálculo del **rendimiento** con el conjunto de **validación**
3. **Repetir** pasos 1 y 2 hasta haber evaluado todas las combinaciones de hiperparámetros
4. Elección de los **hiperparámetros** con los que se obtiene el **mejor rendimiento**

En esta fase debemos evitar obtener un modelo como resultado de un **sobreajuste** (*overfitting*) o **sobregeneralización** (*underfitting*), ya que esto produciría que la salida proporcionada con nuevos datos de entrada tenga un error elevado.



Esquema general del ML: Entrenamiento



Esquema general del ML: Evaluación

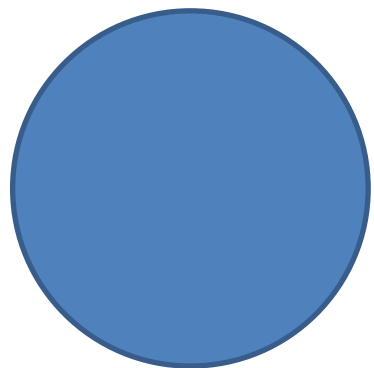
Una vez se tiene un modelo entrenado, es necesario evaluarlo para conocer su rendimiento. Para ver lo bueno o malo que es el modelo generado con el algoritmo de aprendizaje, tenemos una serie de **métricas de evaluación** que nos permitirán ver el nivel de acierto o la tasa de error cometidas en las predicciones bien sea **con los datos de test** (datos que no se han usado para el entrenamiento) o con los datos de entrenamiento (que son datos que el modelo ya ha visto y que por tanto ya conoce y debería predecirlos bien). Las métricas más utilizadas son:

En regresión:

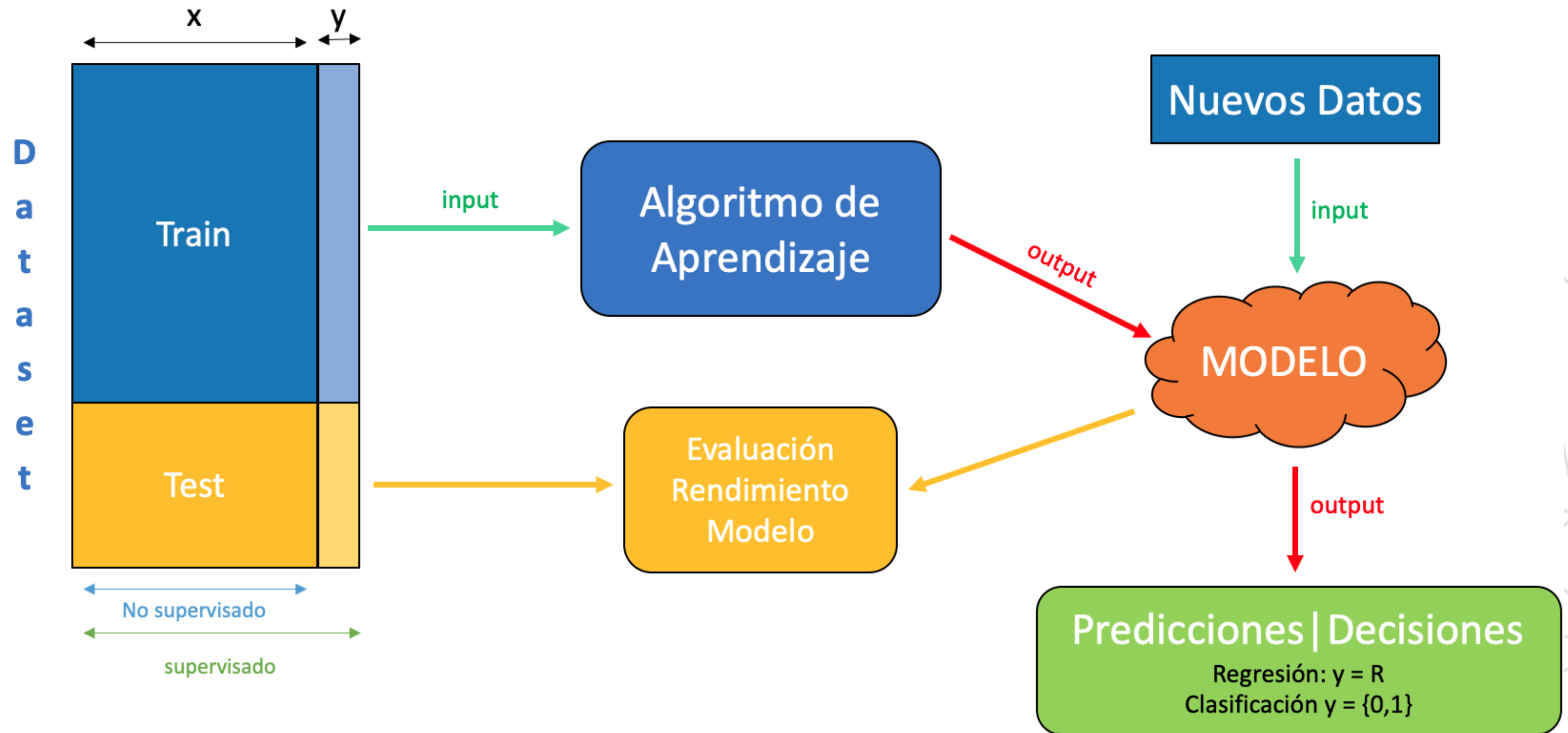
- Error Absoluto Medio (MAE)
- Error Cuadrático Medio (MSE)
- Coeficiente de determinación (R^2)

En clasificación:

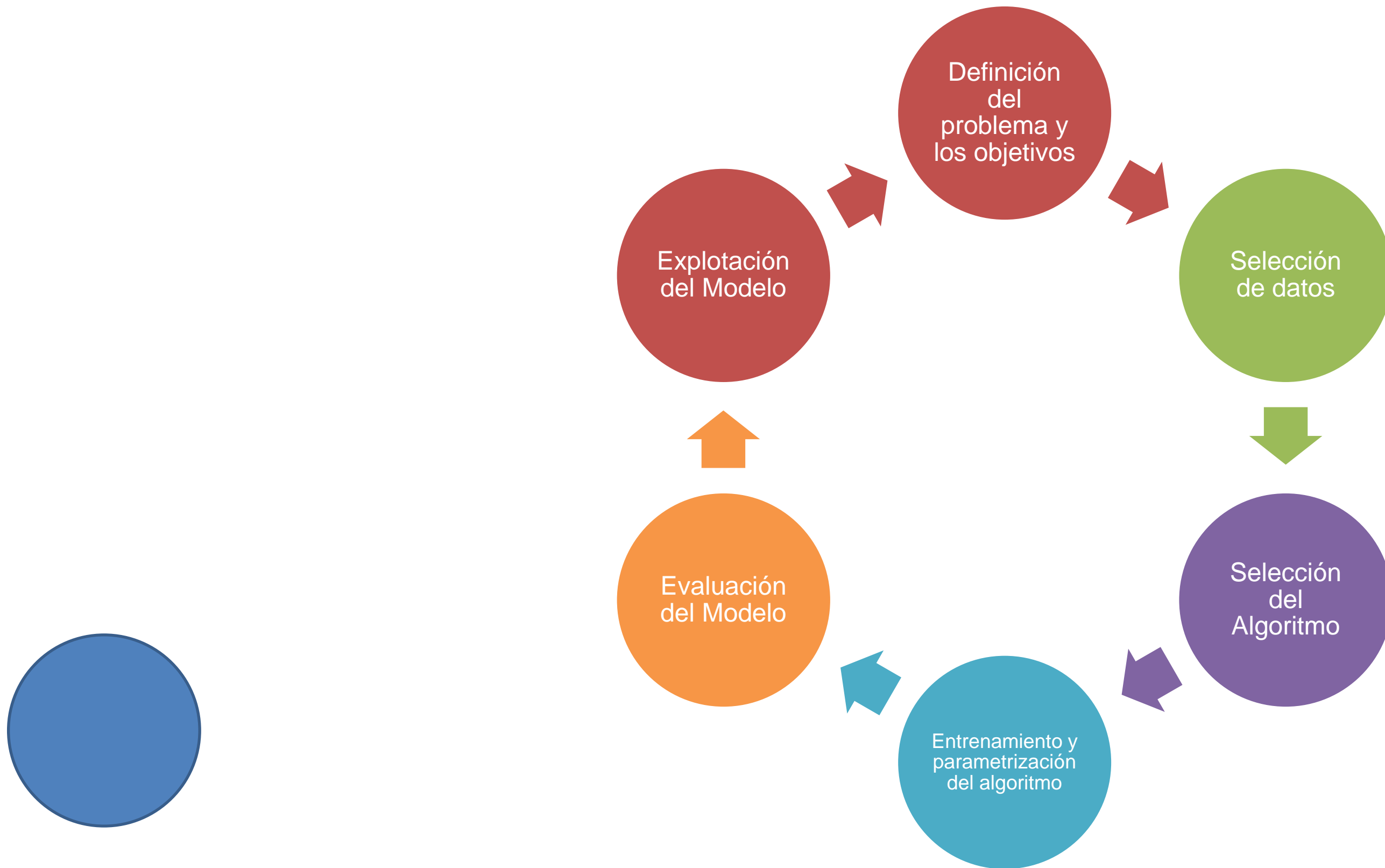
- Matriz de confusión
- Precisión (*Accuracy*)
- Tasa de error (*Error rate*)
- Valor-F (F_1 -score)
- Precisión positiva (PPV)
- Sensibilidad (*Recall* o *TPR*)
- Especificidad (TNR)
- Curva ROC



Metodología del ML: Esquema



Metodología del ML: Resumen



03

Técnicas de selección de datos

Resustitución

Todos los datos se utilizan como test y entrenamiento

Partición

Divide los datos en dos subconjuntos: uno de test y uno de entrenamiento..

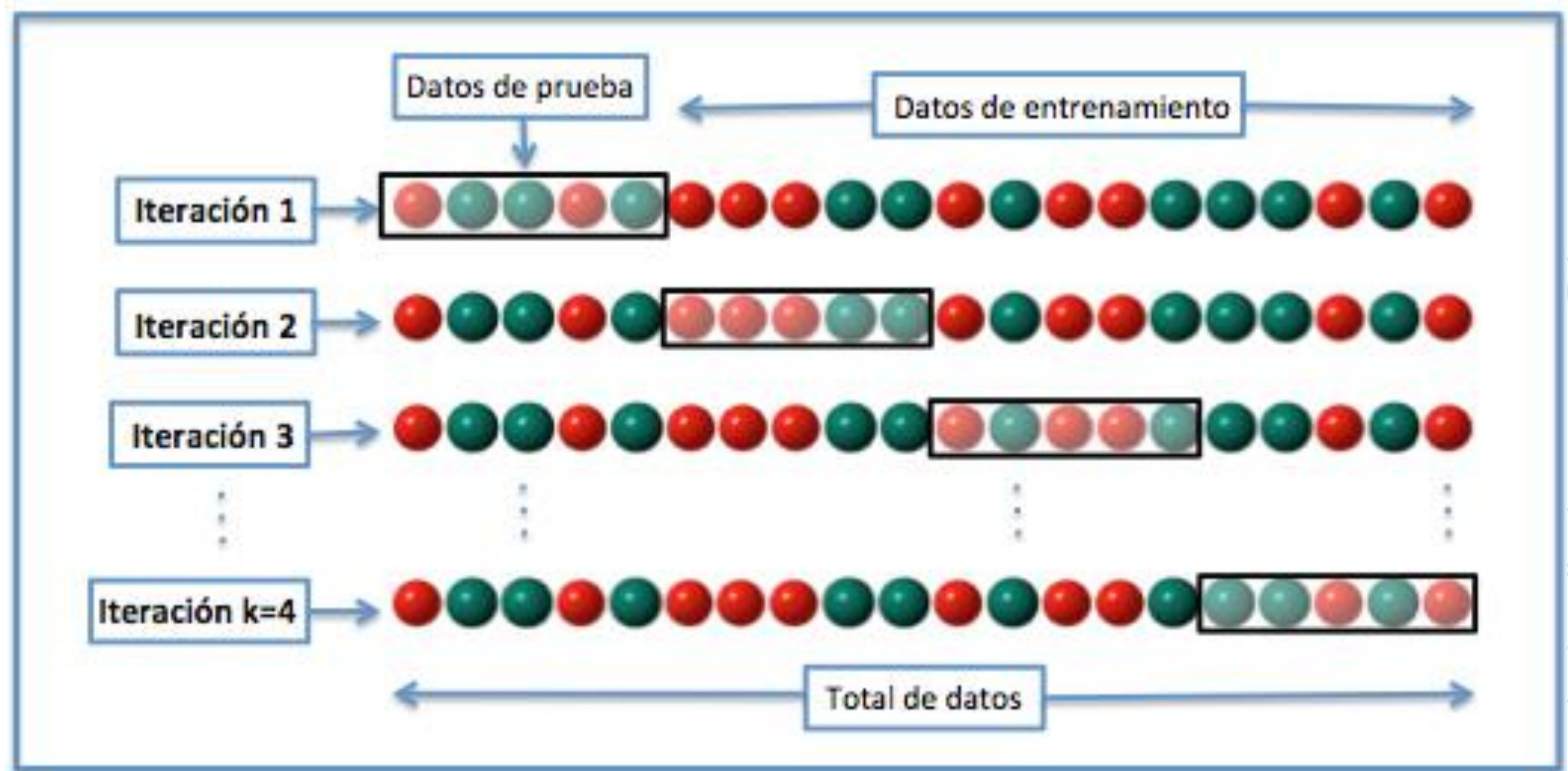
CV

Divide los datos en bloques y cada uno se utiliza como test para un sistema entrenado con el resto de bloques.

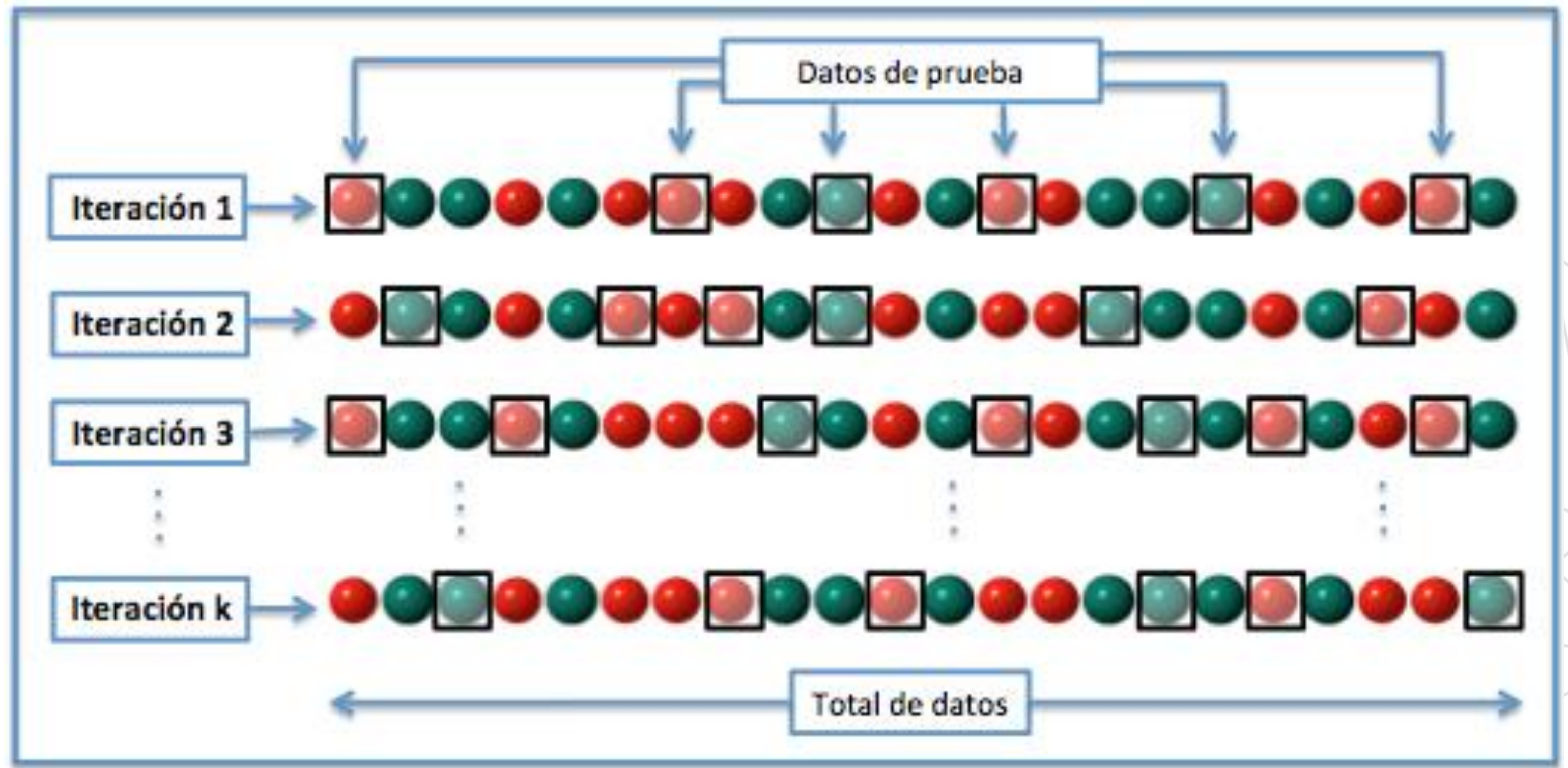
Leave 1 Out

Utiliza cada dato individual como único dato de test.

Validación Cruzada de K iteraciones

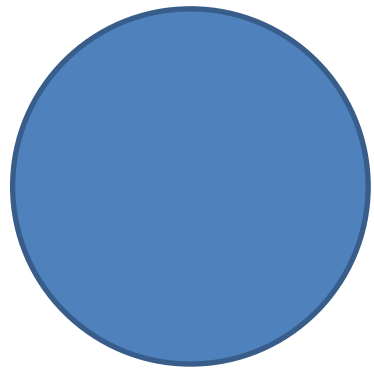


Validación Cruzada aleatoria



04

Scikit-Learn

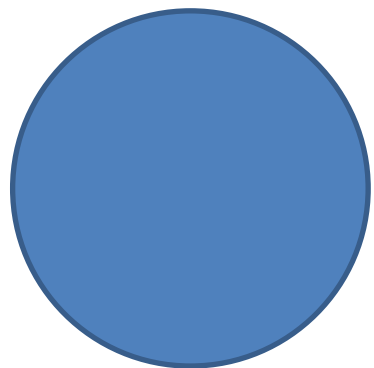


Introducción a Scikit-Learn

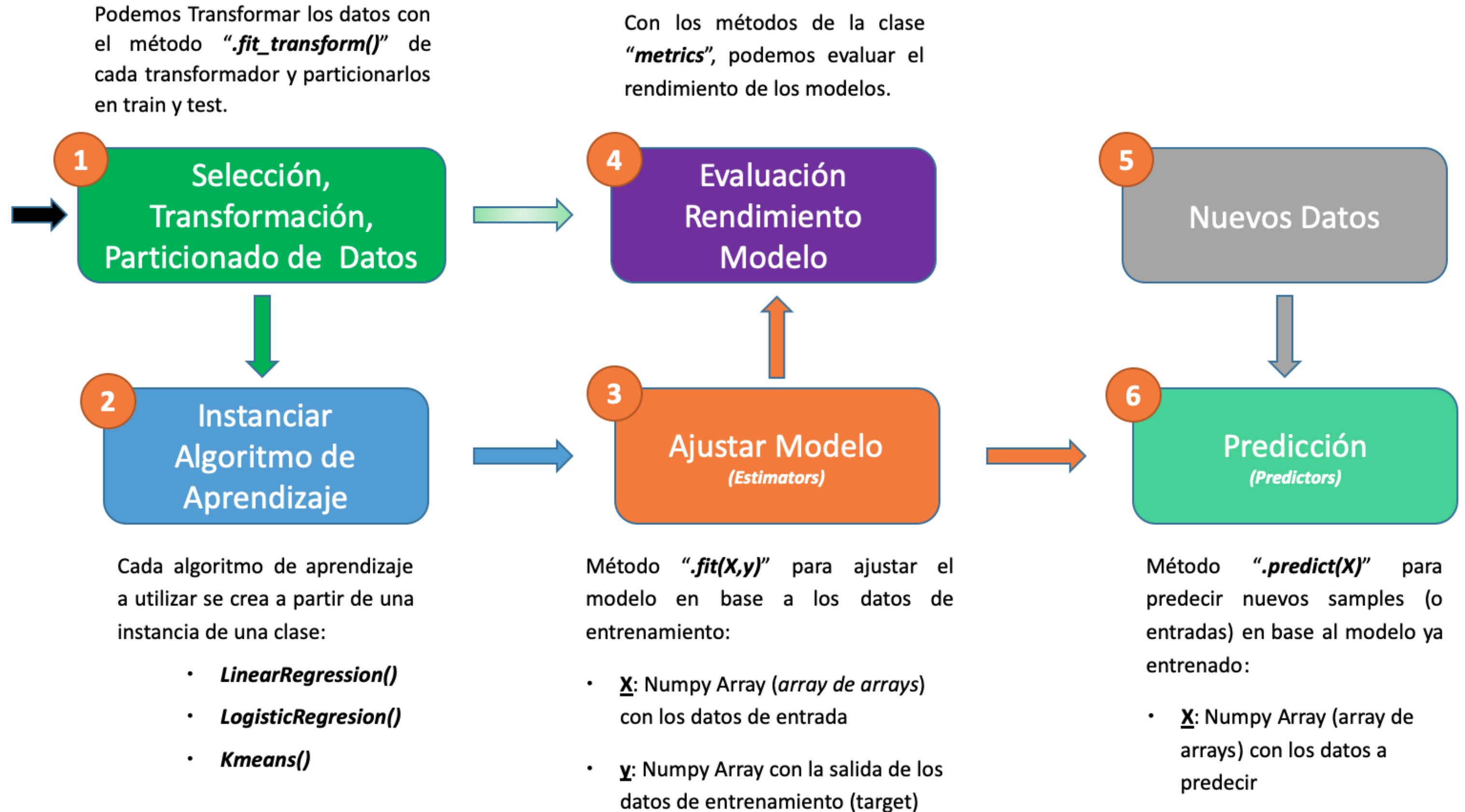
Scikit-learn es una **librería de Machine Learning** gratuita (licencia BSD) creada por David Cournapeau y desarrolla por una comunidad de desarrolladores que cuenta con muchos algoritmos de aprendizaje (regresión, clasificación, clusterización, reducción de la dimensionalidad, etc.), así como otras utilidades para el particionado de datos, generación dummy de datos, evaluación de modelos, etc.

Sus interfaces son:

- **Estimators:** Esta interfaz define el método "*fit()*" para el entrenamiento o ajuste de los modelos.
- **Transformers:** Esta interfaz define los métodos "*transform()*" y "*fit_transform()*" para las clases destinadas a las transformaciones de datos.
- **Predictors:** Esta interfaz define el método "*predict()*" y "*score()*" (para el aprendizaje supervisado) para realizar predicciones y medir los errores cometidos respectivamente.



Flujo de Scikit-Learn



**Muchas gracias
por vuestra
atención.**

Carlos Moreno Morera

Consultor de IA en IBM

carmor06@ucm.es



pontia

