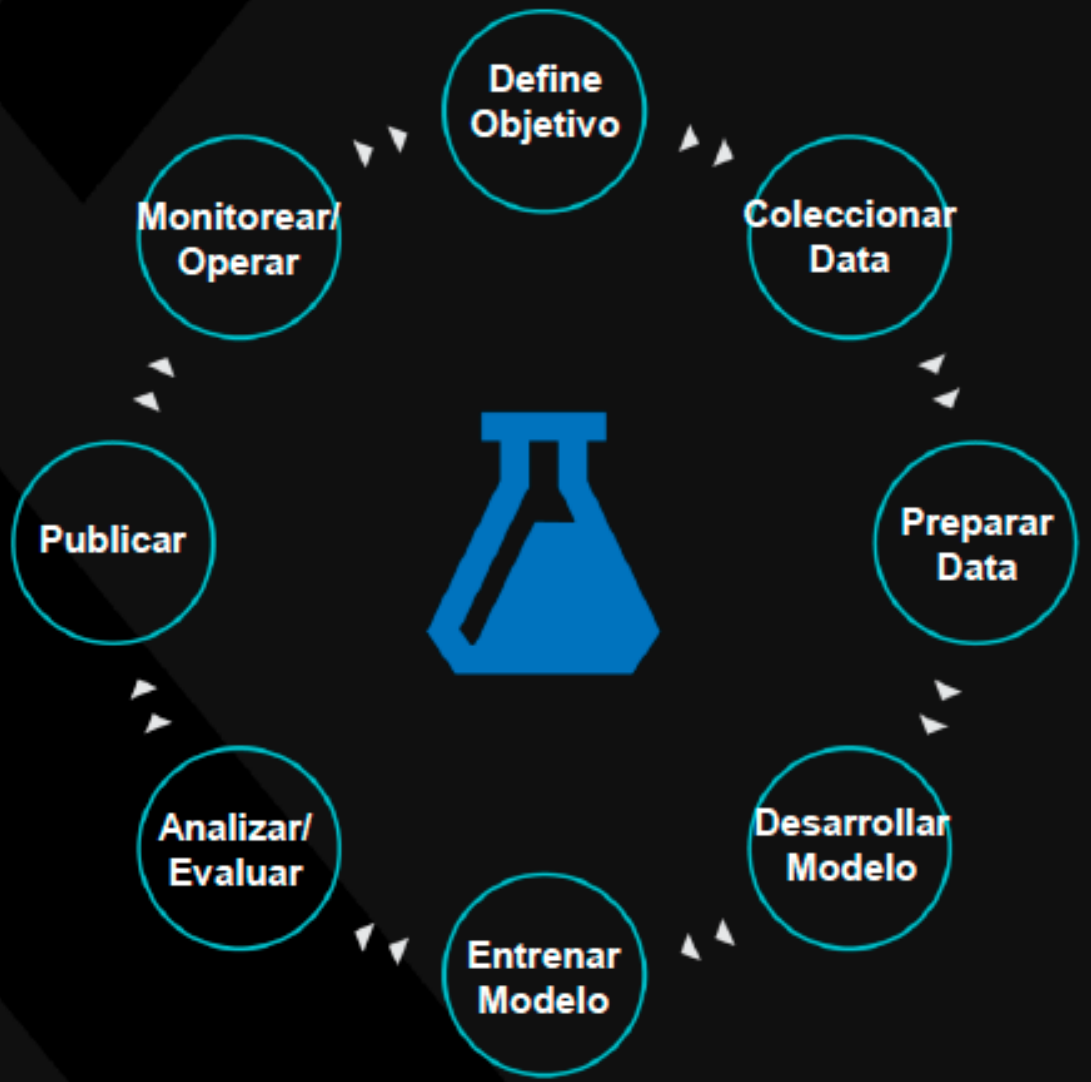
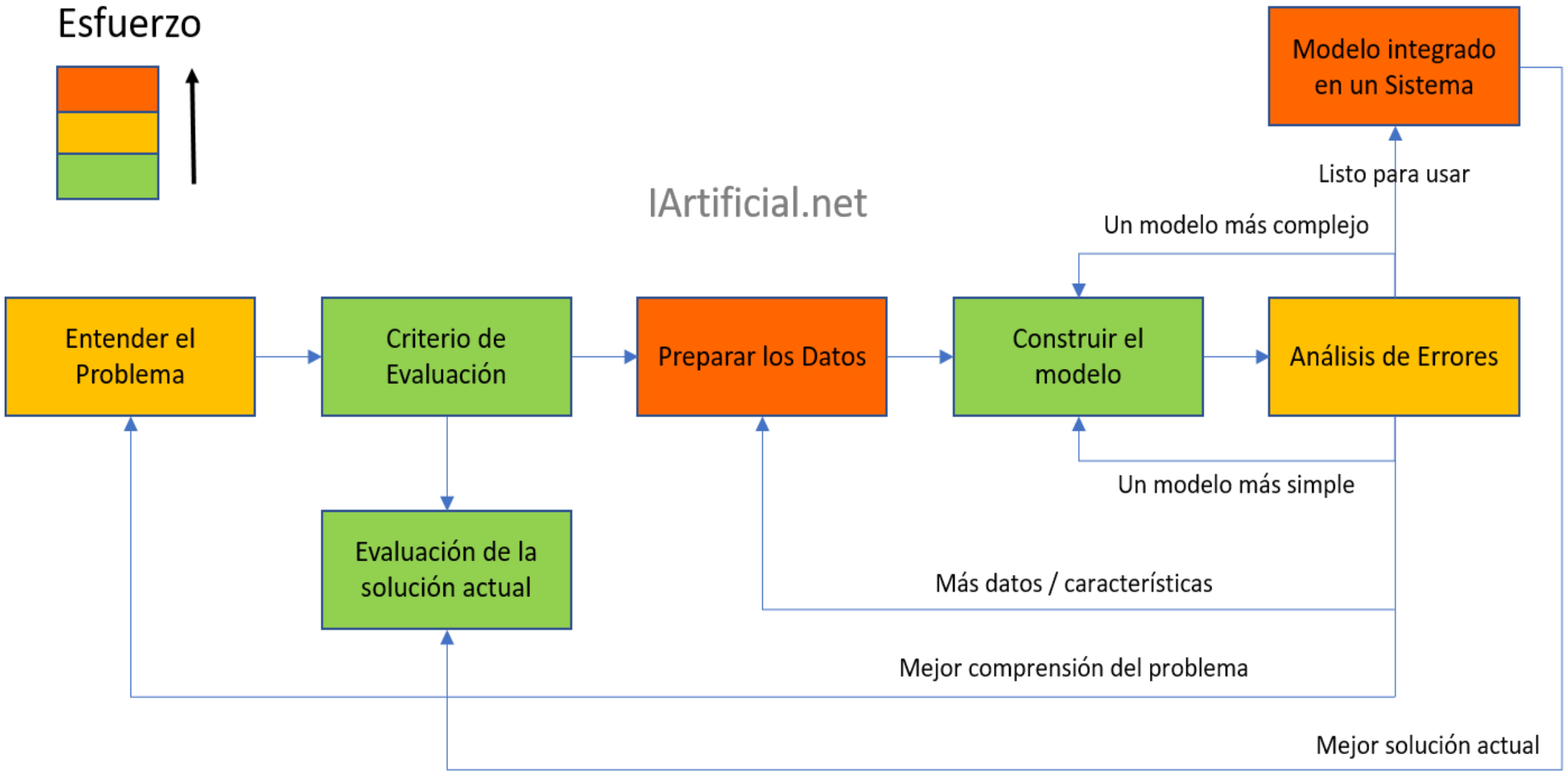
# Flujos de trabajo de Machine Learning

Metodología

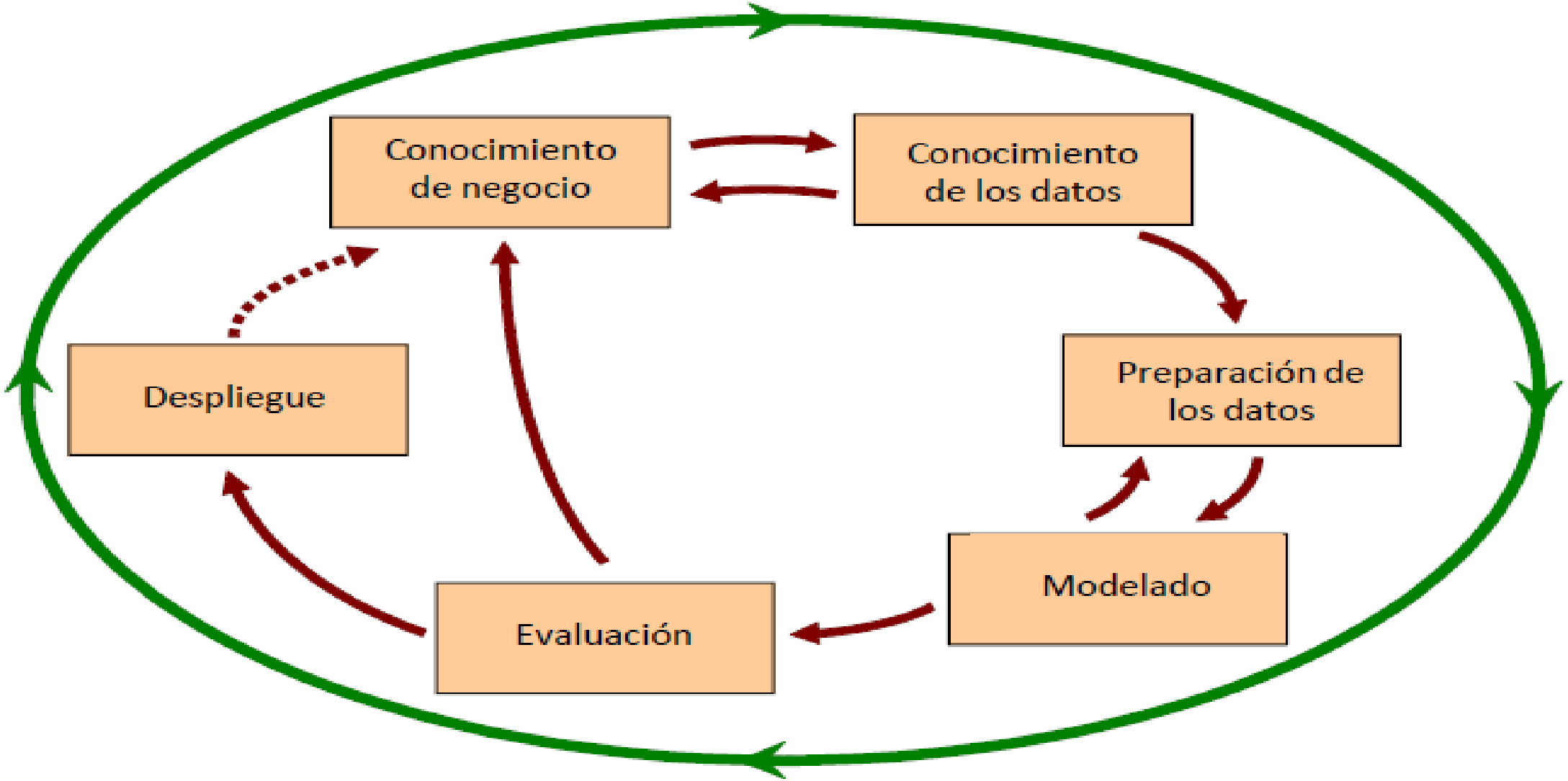
# El Flujo de Trabajo de Machine Learning



# Fases del Proceso de Machine Learning



# Metodología CRISP-DM



# Comprensión de negocio

# Comprensión de negocio

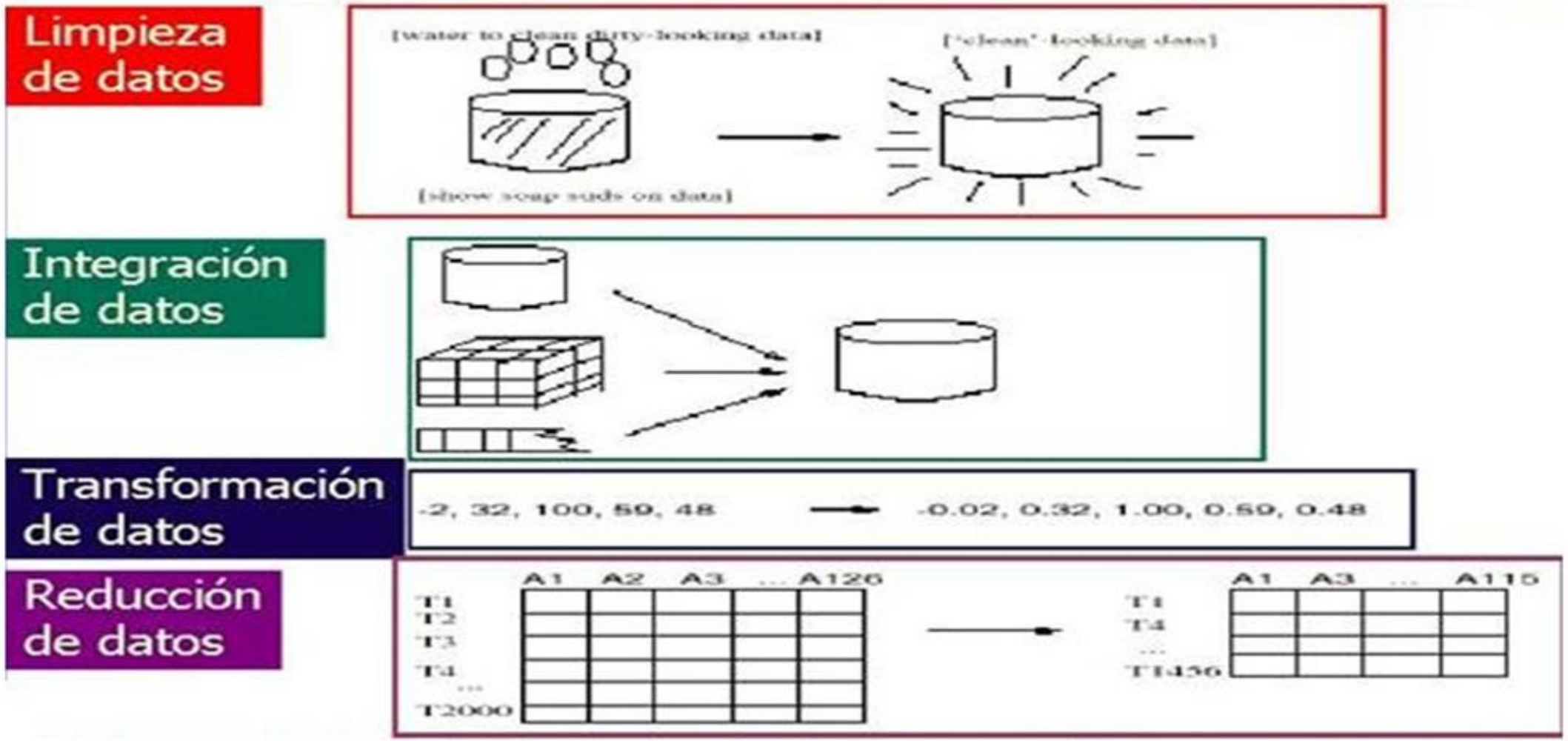
- Comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial:
  - establecimiento de objetivos (contexto inicial, objetivos y criterios de éxito)
  - evaluación de la situación (recursos, requerimientos, suposiciones, restricciones, riesgos y contingencias, terminología y costes y beneficios)
  - establecimiento de los objetivos de la minería de datos (planteamiento de los objetivos y criterios de éxito de la minería de datos)
  - creación del plan de proyecto (plan de proyecto y evaluación inicial de herramientas y técnicas).

# Comprensión de los datos

# Comprensión de los datos

- Recopilar y familiarizarse con los datos, identificar los problemas de calidad y discernir datos potenciales o subconjuntos que pueden ser interesantes de analizar (de acuerdo con los objetivos empresariales de la fase anterior):
  - Recopilación inicial de datos (informe de recopilación),
  - Descripción de los datos (informe de descripción),
  - Exploración de los datos (informe de exploración)
  - Verificación de la calidad de los datos (informe de calidad).

# Preparación de los datos

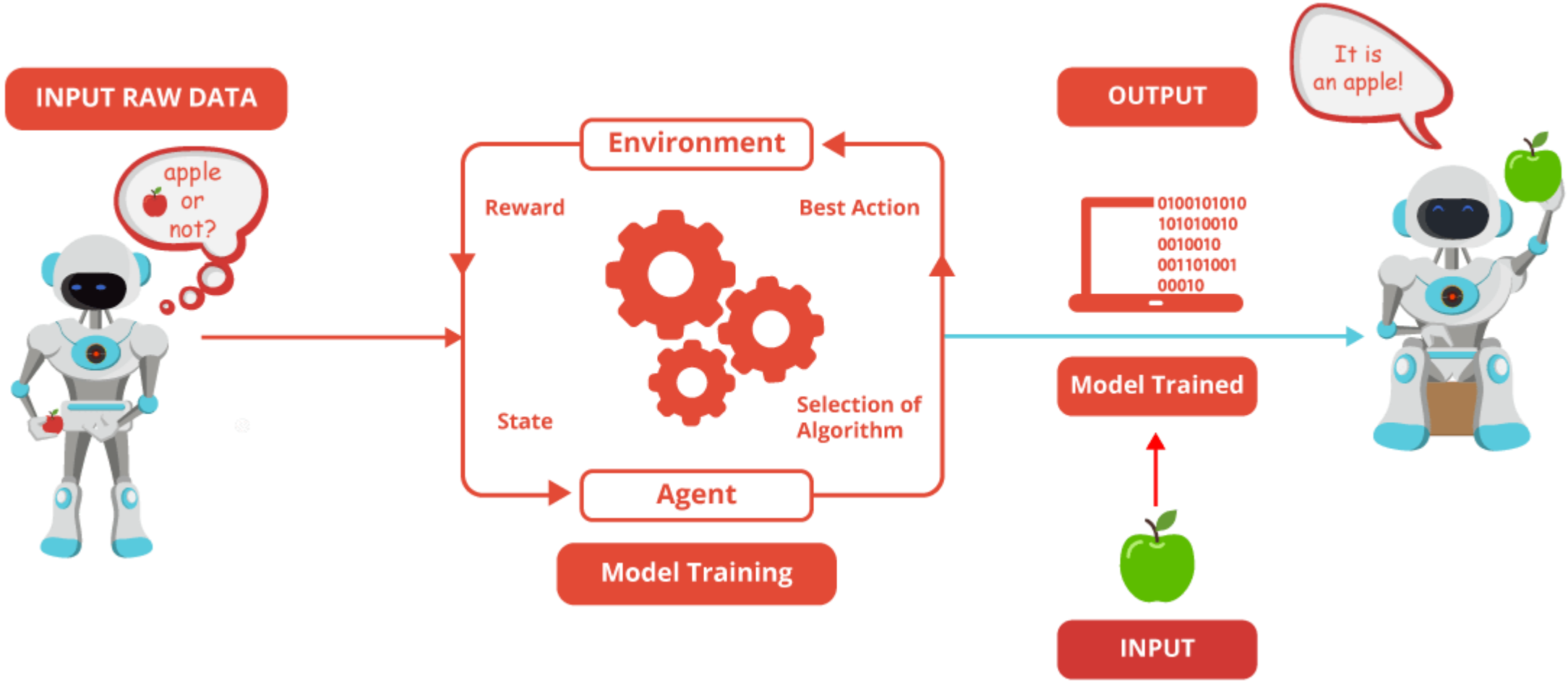


# Preparación de los datos

El objetivo de esta fase es obtener la “vista minable”. Aquí encontramos: integración, selección, limpieza y transformación.

- selección de datos (motivos de inclusión/exclusión),
- limpieza de datos (informe de limpieza de datos),
- construcción de datos (atributos derivados, archivos generados),
- integración de datos (mezcla de datos)
- formateo de datos (datos reformados).

# Modelado de los datos



# Modelado de los datos

- Es la aplicación de técnicas de modelado o minería de datos a las vistas minables anteriores. Subfases:
  - selección de la técnica de modelado (técnica de modelado, suposición de modelado),
  - diseño de evaluación (diseño de prueba),
  - construcción del modelo (parámetros elegidos, modelos, descripción del modelo)
  - evaluación del modelo (medidas del modelo, revisión de los parámetros escogidos).

# Evaluación del modelo

# Evaluación del modelo

Es necesario evaluar (desde el punto de vista de la meta) los modelos de la fase anterior. En otras palabras, si el modelo es útil para responder algunos de los requisitos comerciales.

- evaluación del resultado (evaluación de los resultados de la minería de datos, modelos aprobados),
- revisar el proceso (proceso de revisión)
- establecimiento de los siguientes pasos (lista de posibles acciones, decisiones).

# Despliegue

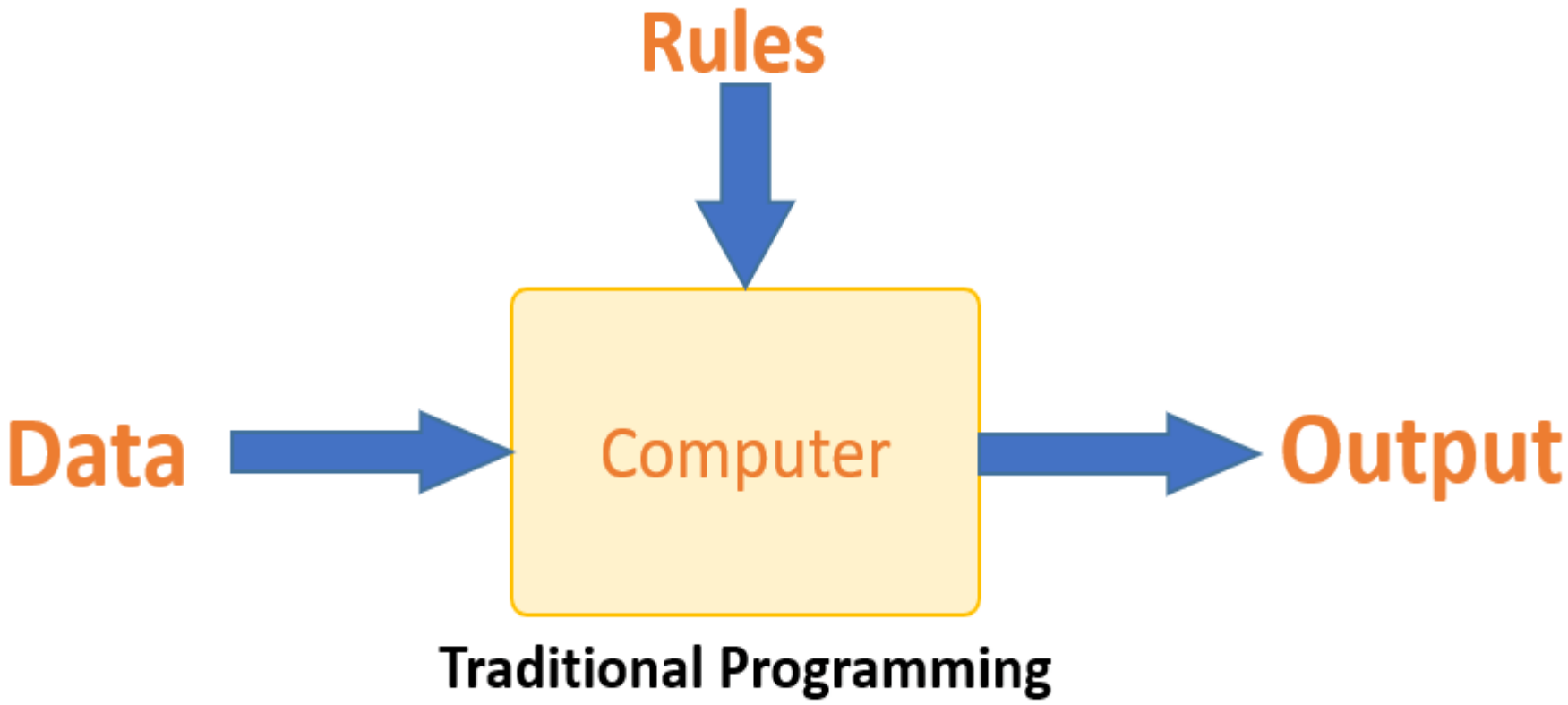
# Despliegue

- La idea es explotar el potencias de los modelos extraídos, integrarlos en los procesos de toma de decisiones de la organización, repartir informes sobre el conocimiento extraído
  - planificación del despliegue (plan de despliegue),
  - planificación del mantenimiento y monitorización(plan de monitorización y mantenimiento),
  - creación del informe final (informe final, presentación final),
  - revisión del proyecto(documentación de la experiencia).

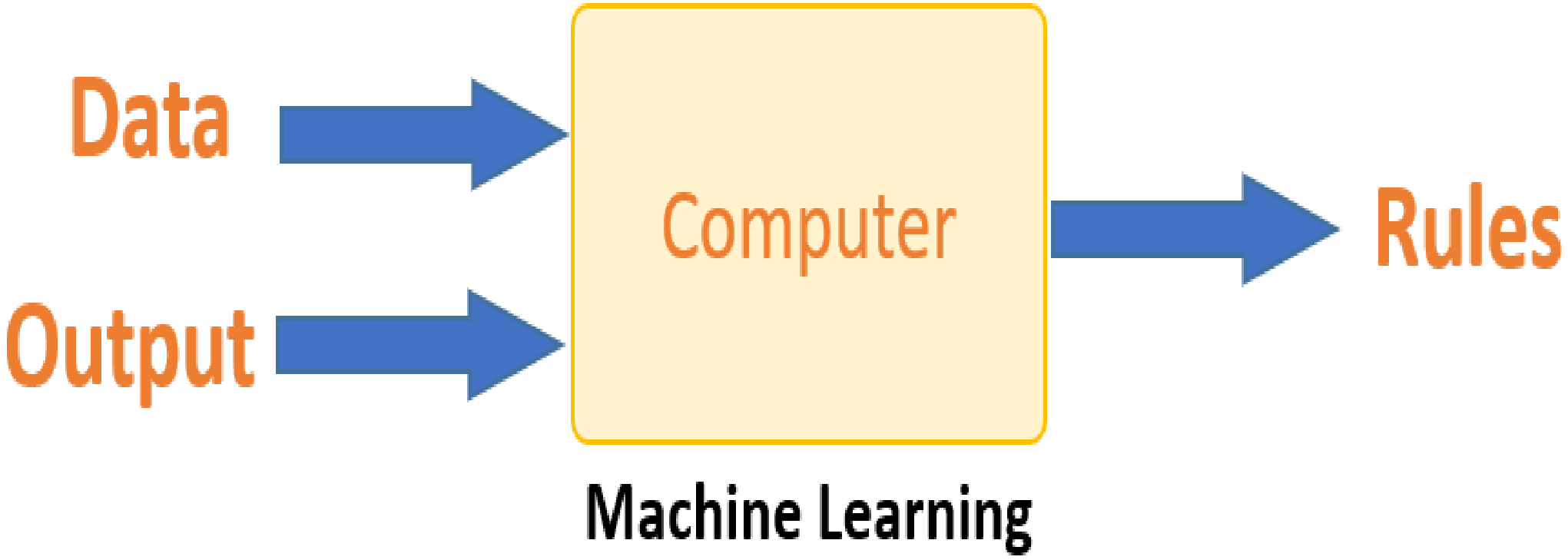
# Machine Learning

*“Machine Learning es la ciencia que permite que las computadoras aprendan y actúen como lo hacen los humanos, mejorando su aprendizaje a lo largo del tiempo de una forma autónoma, alimentándolas con datos e información en forma de observaciones e interacciones con el mundo real.”* — ***Dan Fagella***

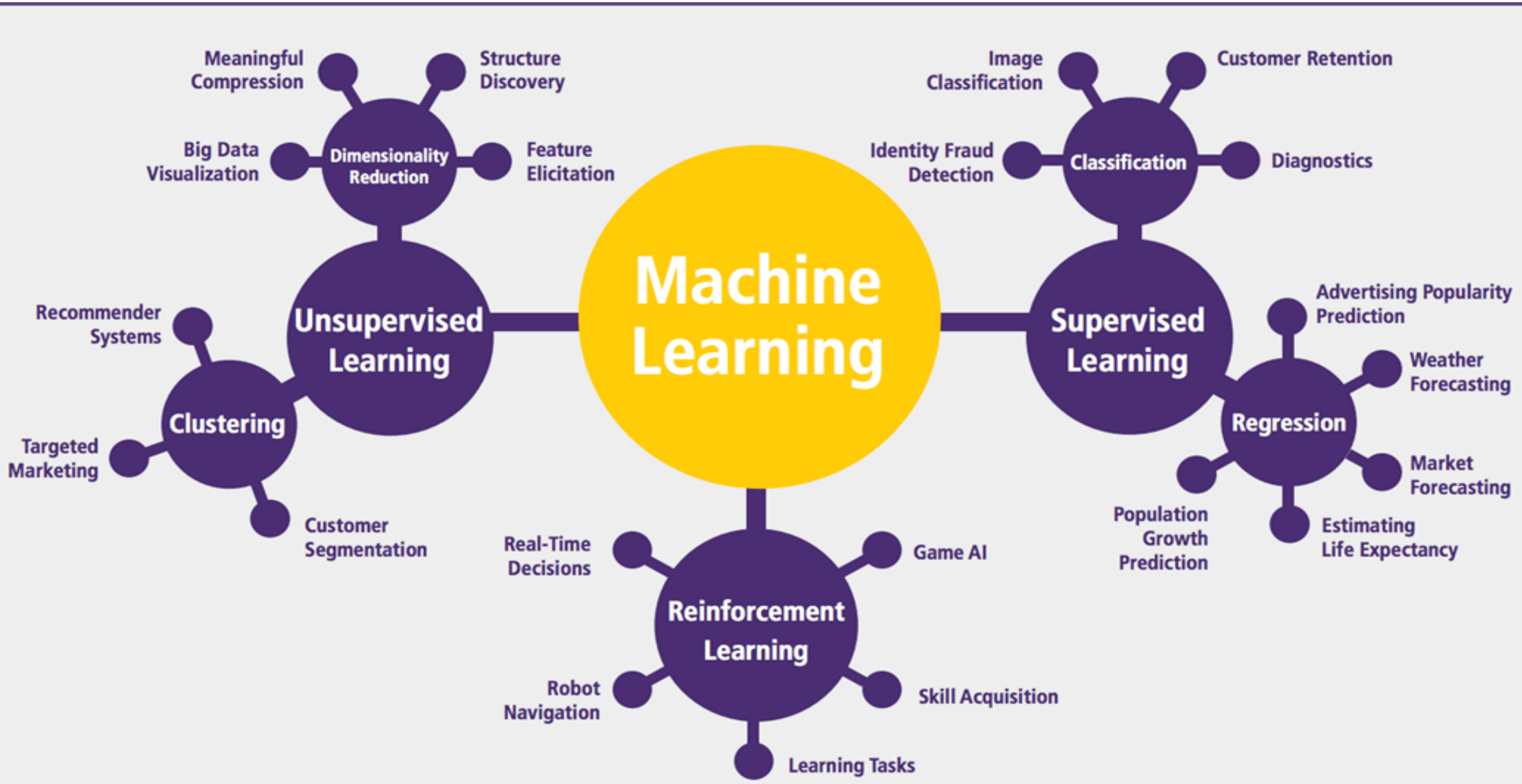
# Machine Learning



# Machine Learning



# Machine Learning



# Machine Learning

•**Banca:** Un ejemplo práctico del Machine Learning en banca sería, por ejemplo, identificar clientes con **perfiles de alto riesgo** o bien analizar las transacciones para **detectar signos de fraude**.

•**Gobiernos:** La administración tiene múltiples fuentes de datos de las que se pueden **extraer información**: tráfico, padrón de habitantes, patrones de conductas, etc.

•**Sanidad:** Los sensores que las personas pueden llevar, analizan en tiempo real su estado de salud y la tecnología puede ayudar a expertos médicos a analizar esos datos para identificar tendencias o alertas que permitan **realizar diagnósticos mejorados**.

# Machine Learning

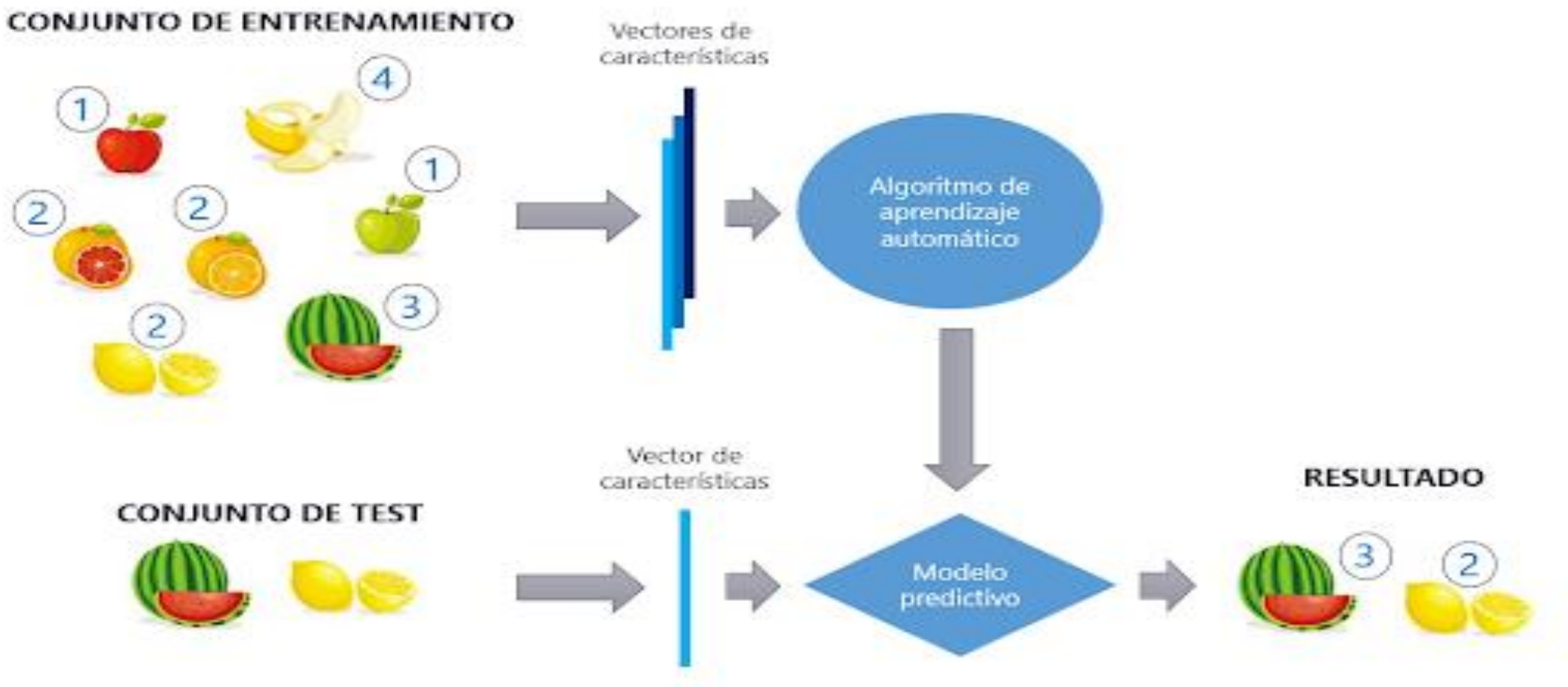
- **Marketing y ventas:** Se utiliza en las páginas web, principalmente en el comercio electrónico. Se capturan los datos del usuario, se analiza su comportamiento y se utilizan para **personalizar una experiencia de compra**. Ese es ya el presente del comercio electrónico.

- **Transporte:** Se analizan los datos de tráfico disponibles para identificar patrones y tendencias. Con ese objetivo, las empresas de transporte pueden hacer rutas más eficientes y anticipar problemas para incrementar la rentabilidad.

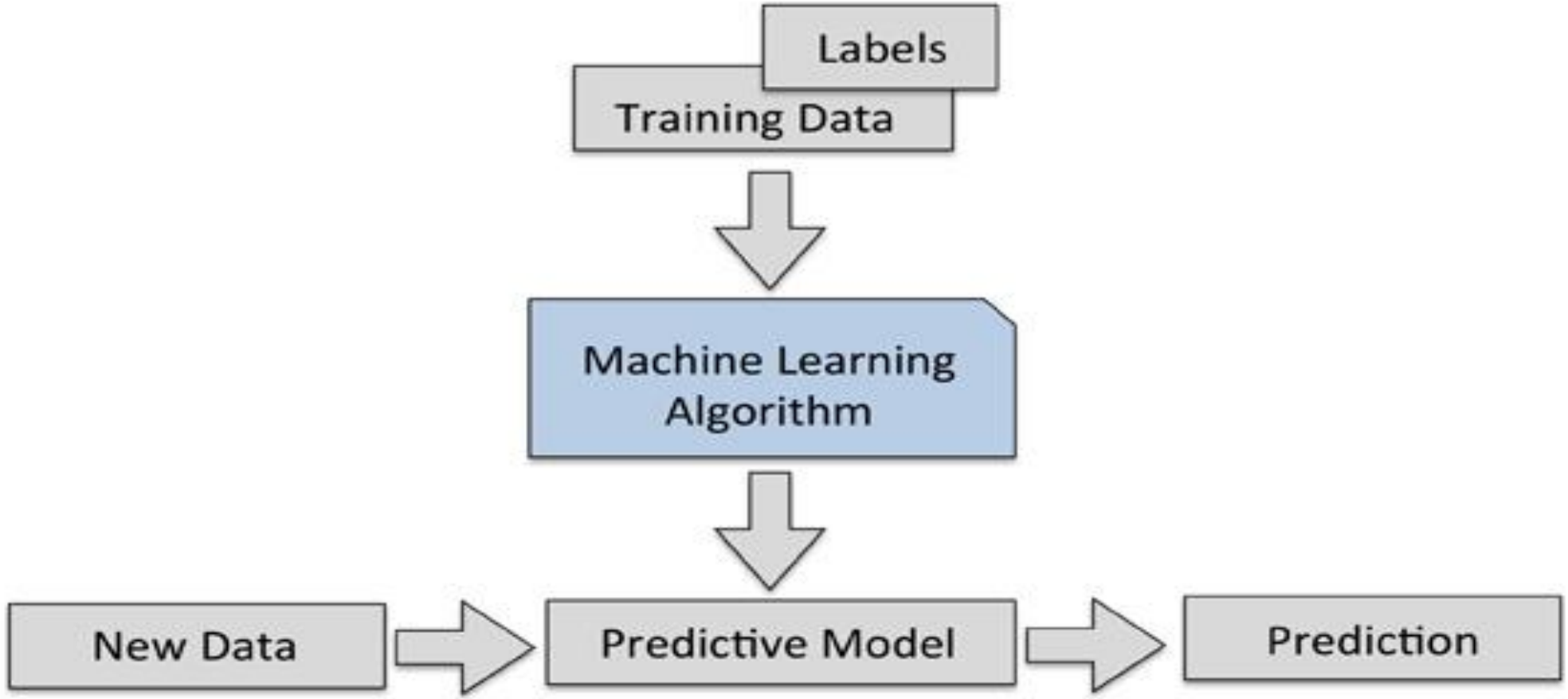
# Aprendizaje Supervisado

- Se refiere a un tipo de modelos de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos.
- Los modelos aprenden de esos resultados conocidos y realizan ajustes en sus parámetros interiores para adaptarse a los datos de entrada.
- Una vez el modelo es entrenado adecuadamente, y los parámetros internos son coherentes con los datos de entrada y los resultados de la batería de datos de entrenamiento, el modelo podrá realizar predicciones adecuadas ante nuevos datos no procesados previamente.

# Aprendizaje Supervisado



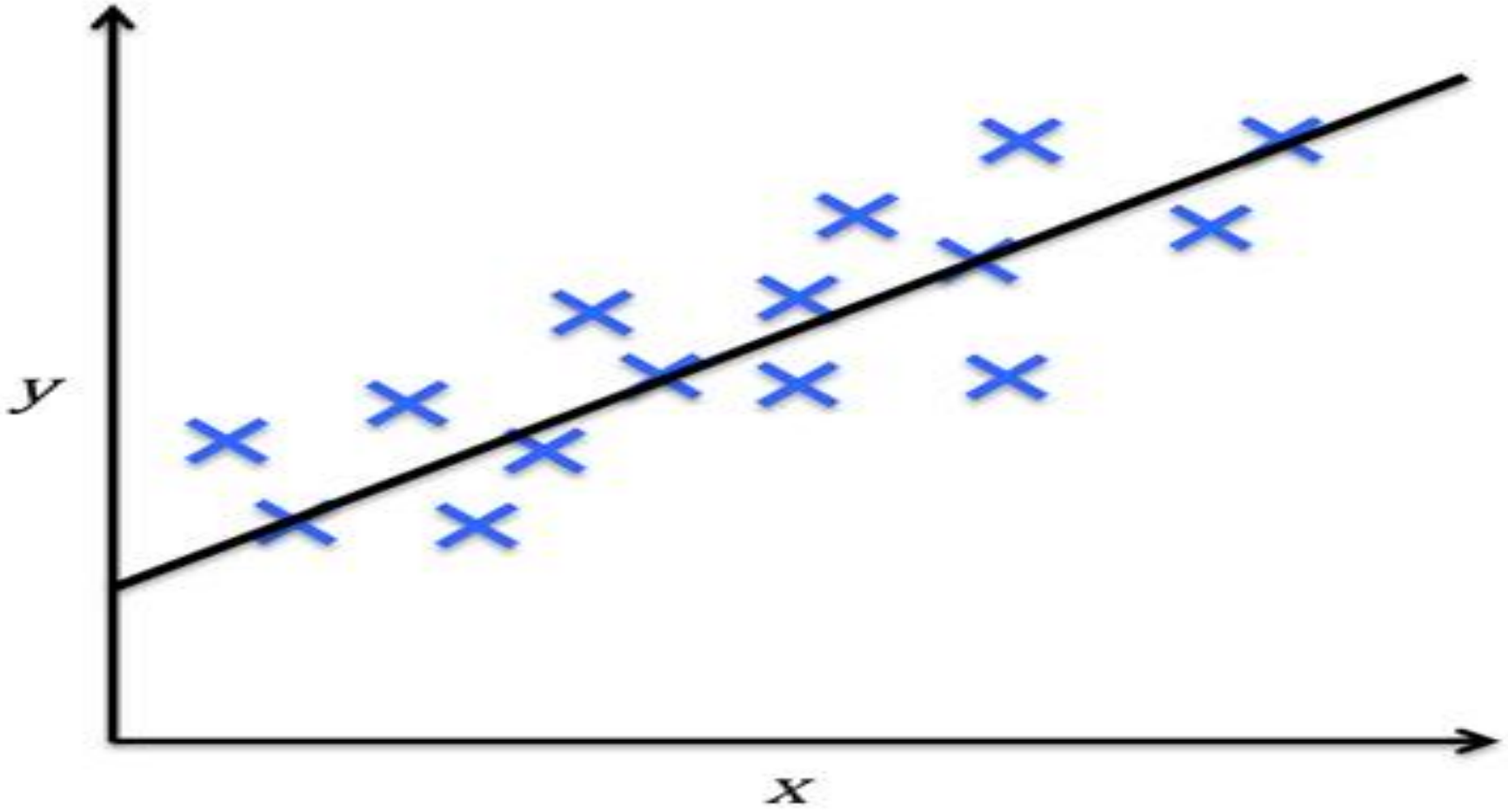
# Aprendizaje Supervisado



# Aplicaciones principales de aprendizaje supervisado

- La **regresión** se utiliza para asignar categorías a datos sin etiquetar.
- En este tipo de aprendizaje tenemos un número de variables predictoras (explicativas) y una variable de respuesta continua (resultado), y se tratará de encontrar una relación entre dichas variables que nos proporciones un resultado continuo.

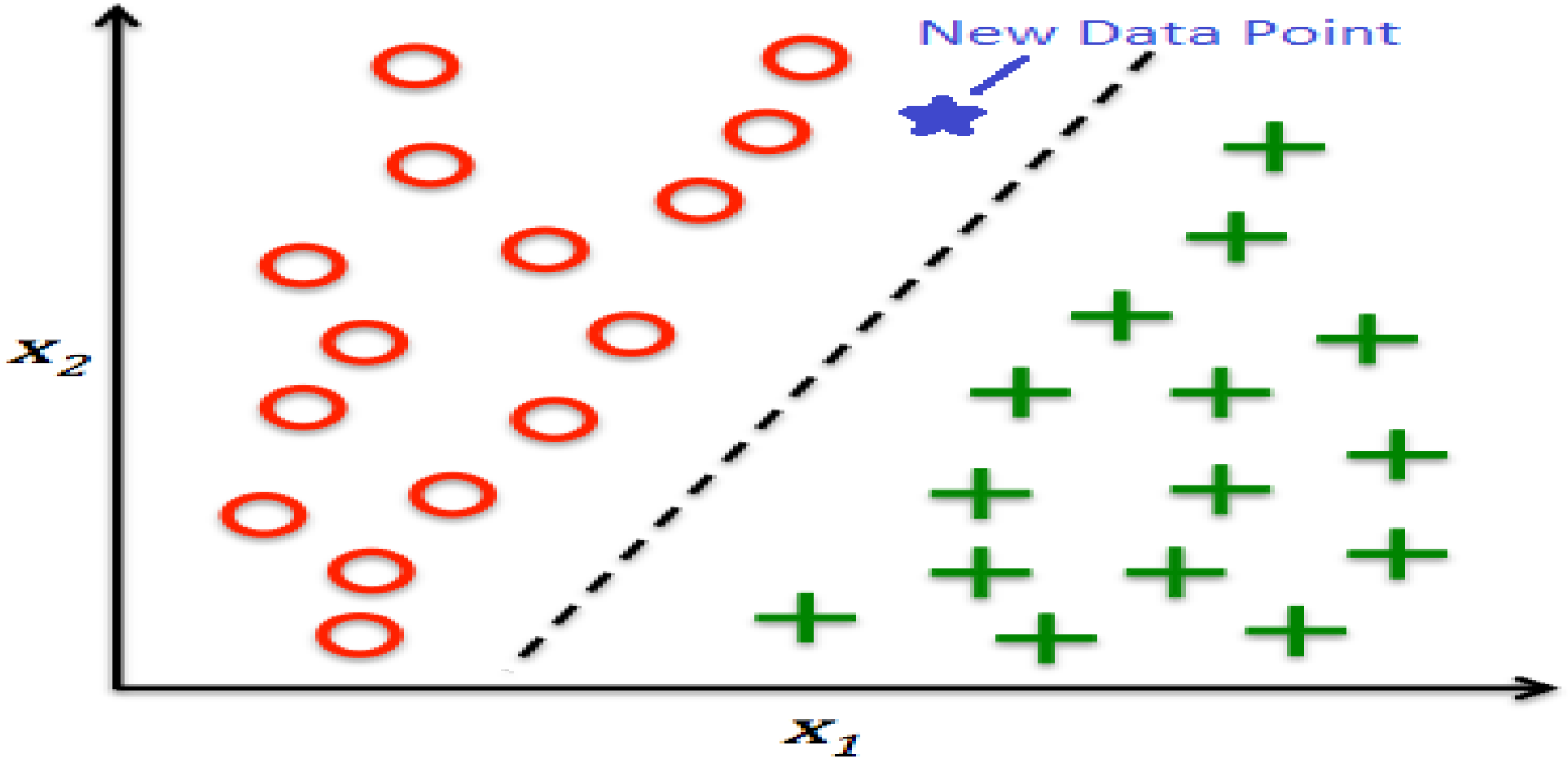
# Aplicaciones principales de aprendizaje supervisado



# Aplicaciones principales de aprendizaje supervisado

- **Clasificación** es una sub-categoría de aprendizaje supervisado en la que el objetivo es predecir las clases categóricas (valores discretos, no ordenados, pertenencia a grupos).
- El ejemplo típico es la detección de correo spam, que es una clasificación binaria (un email es spam — valor “1”- o no lo es — valor “0” -)

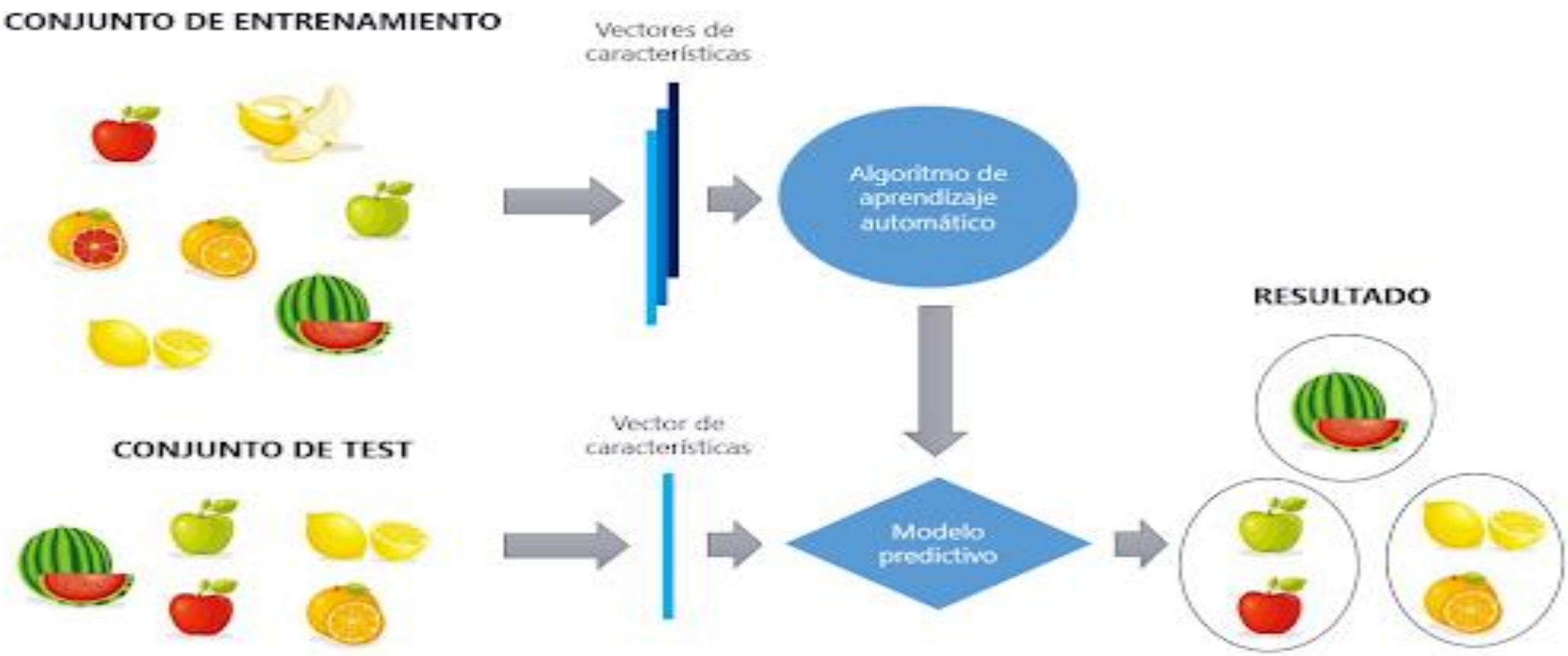
# Aplicaciones principales de aprendizaje supervisado



# Aprendizaje No Supervisado

- En el aprendizaje no supervisado, trataremos con datos sin etiquetar cuya estructura es desconocida. El objetivo será la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar.

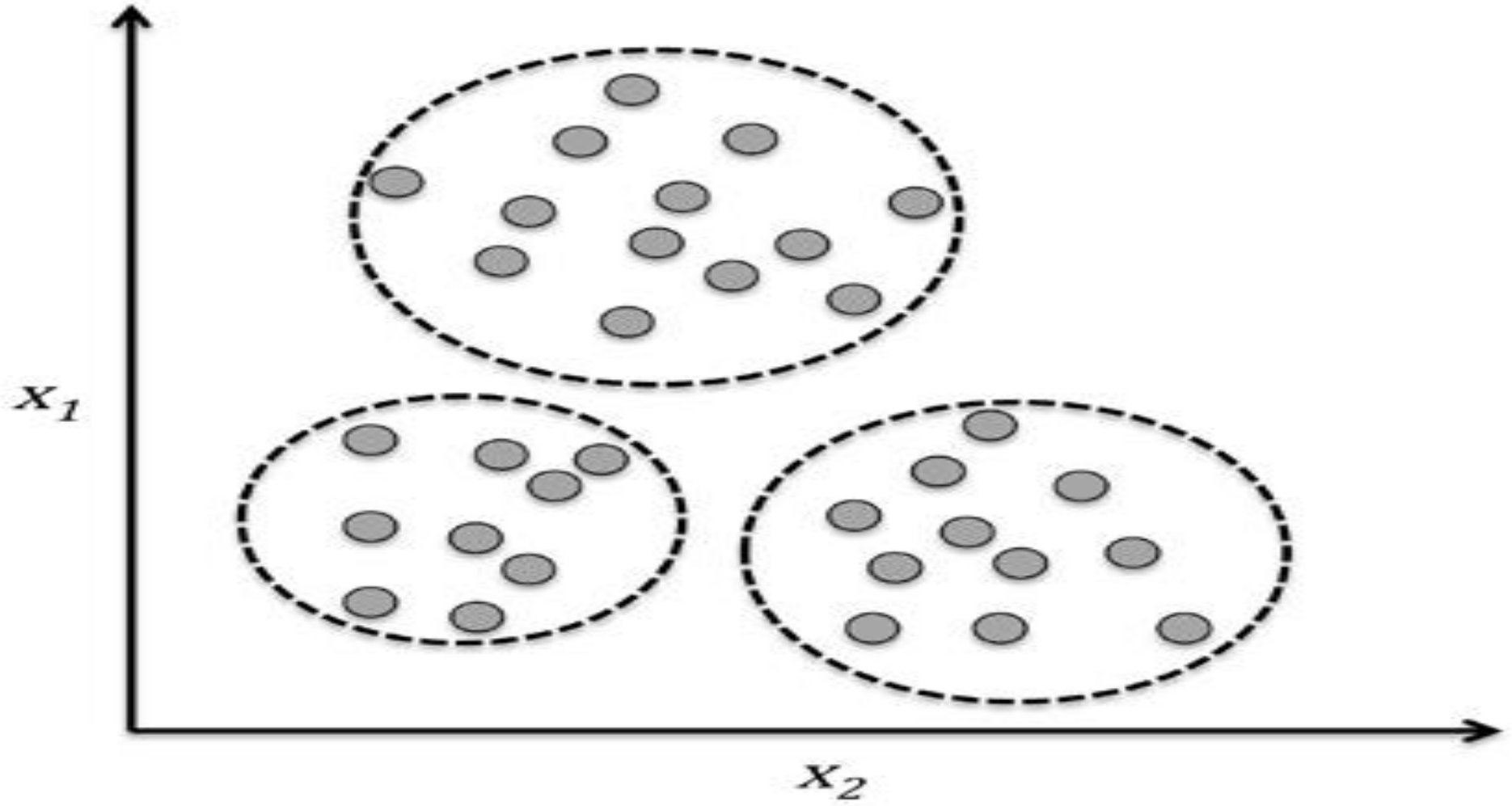
# Aprendizaje No Supervisado



# Aprendizaje No Supervisado

- El **agrupamiento** es una técnica exploratoria de análisis de datos, que se usa para organizar información en grupos con significado sin tener conocimiento previo de su estructura.
- Cada grupo es un conjunto de objetos similares que se diferencia de los objetos de otros grupos.
- El objetivo es obtener un numero de grupos de características similares.

# Aprendizaje No Supervisado



# Aprendizaje No Supervisado

- Es común trabajar con datos en los que cada observación se presenta con alto número de características, en otras palabras, que tienen alta dimensionalidad.
- Este hecho es un reto para la capacidad de procesamiento y el rendimiento computacional de los algoritmos de Machine Learning.
- La reducción dimensional es una de las técnicas usadas para mitigar este efecto.
- La reducción dimensional funciona encontrando correlaciones entre las características, lo que implica que existe información redundante, ya que alguna característica puede explicarse parcialmente con otras (por ejemplo, puede existir dependencia lineal).
- Estas técnicas eliminan “ruido” de los datos (que puede también empeorar el comportamiento del modelo), y comprimen los datos en un sub-espacio más reducido, al tiempo que retienen la mayoría de la información relevante.

# Aprendizaje No Supervisado