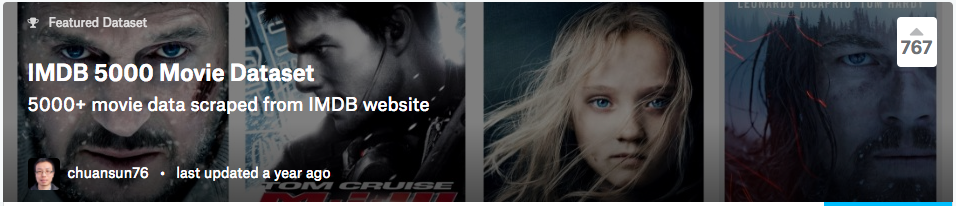
**Curso de Machine Learning Aplicado con scikit-learn**

*Juan Pablo Morales, CTO and Founder, Arara*





# Modelo CRISP-DM Image result for crisp-dm

# ¿Por qué es tan importante definir el problema?

Errores comúnes:

* *No tienes problemas por resolver*
* *Existían soluciones más simples*
* *No puedes medir el impacto de tu modelo*
* *No sabes si el problema ya ha sido resuelto*
* *El problema era imposible de resolver*

-> Definir **UN PROBLEMA REAL** de tu empresa o aplicación, de forma clara y precisa, para saber si corresponde atacarlo con Machine Learning

# Preguntas clave por responder

## Reconoce el tipo de aprendizaje que necesitas

1. ¿Qué beneficio piensas generar y para quién?
2. Cuál de las siguientes funcionalidades te sería más útil para lograr ese objetivo:
   1. Predecir alguna métrica
   2. Predecir una etiqueta
   3. Agrupar elementos similares
   4. Optimizar un proceso con prueba y error

Los dos primeros corresponden a **Aprendizaje supervisado**. Nos enfocaremos en este ya que es el que está teniendo más impacto en la industria (cf. Andrew NG).

***Aterriza tu problema de aprendizaje supervisado***

1. ¿Lo que quieres predecir es un valor de qué tipo?
   1. Continuo
   2. Discreto
2. ¿Cuál es tu definición de éxito de una predicción?
3. ¿Con qué datos contarías para hacer esta predicción?
4. ¿La pregunta que estás tratando de resolver pertenece a alguna disciplina en particular?
5. ¿Considerando tu intuición en la disciplina, crees que los datos te permitan predecir tu objetivo?

**Nuestro problema: Predicción de Ingresos de películas.**

Contexto: *Somos un ente gubernamental que quiere definir sus políticas de financiamiento de producciones cinematográficas nacionales.*

***Reconoce el tipo de aprendizaje que necesitas***

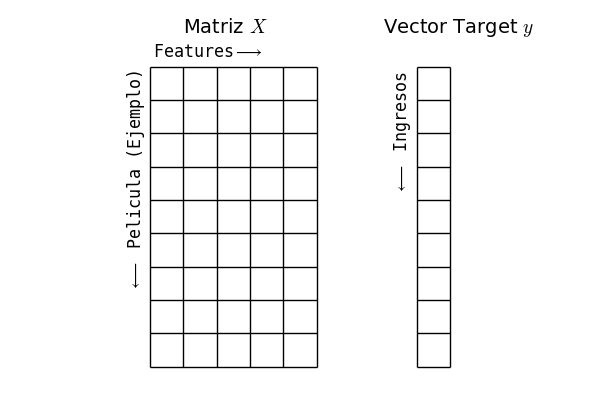
1. **Ayudar** a la producción de películas de calidad **que no logran ser autosustentables**.
2. Nos sería útil saber que películas tienen más dificultad para recuperar en sus presupuestos. Por consiguiente, queremos **predecir una métrica**: el ingreso mundial generado por una película.

***Aterriza tu problema de aprendizaje supervisado***

1. Los ingresos de una película corresponden a **valores continuos**.
2. Mi éxito será **"qué tan cerca estoy del valor real de ingreso generado por la pelicula"**.
3. Me basaré en **bases de datos públicas de internet**
4. El dominio de trabajo es la **industria del cine**, en particular de la distribución de peliculas.
5. Sí, de forma general existen bastantes caracteristicas que me pueden ayudar a saber que película será exitosa como: **calidad, actores, presupuesto, etc...**

* *Quiero predecir ingresos de péliculas, para tomar mejores decisiones de financiamiento, con base a una regresión sobre datos de películas extraídos de internet.*
* *Mi evaluación del éxito será la precisión de mis predicciones.*
* *Podré apoyarme en conocimientos específicos de la industria.*

# Un poco de terminología de ML



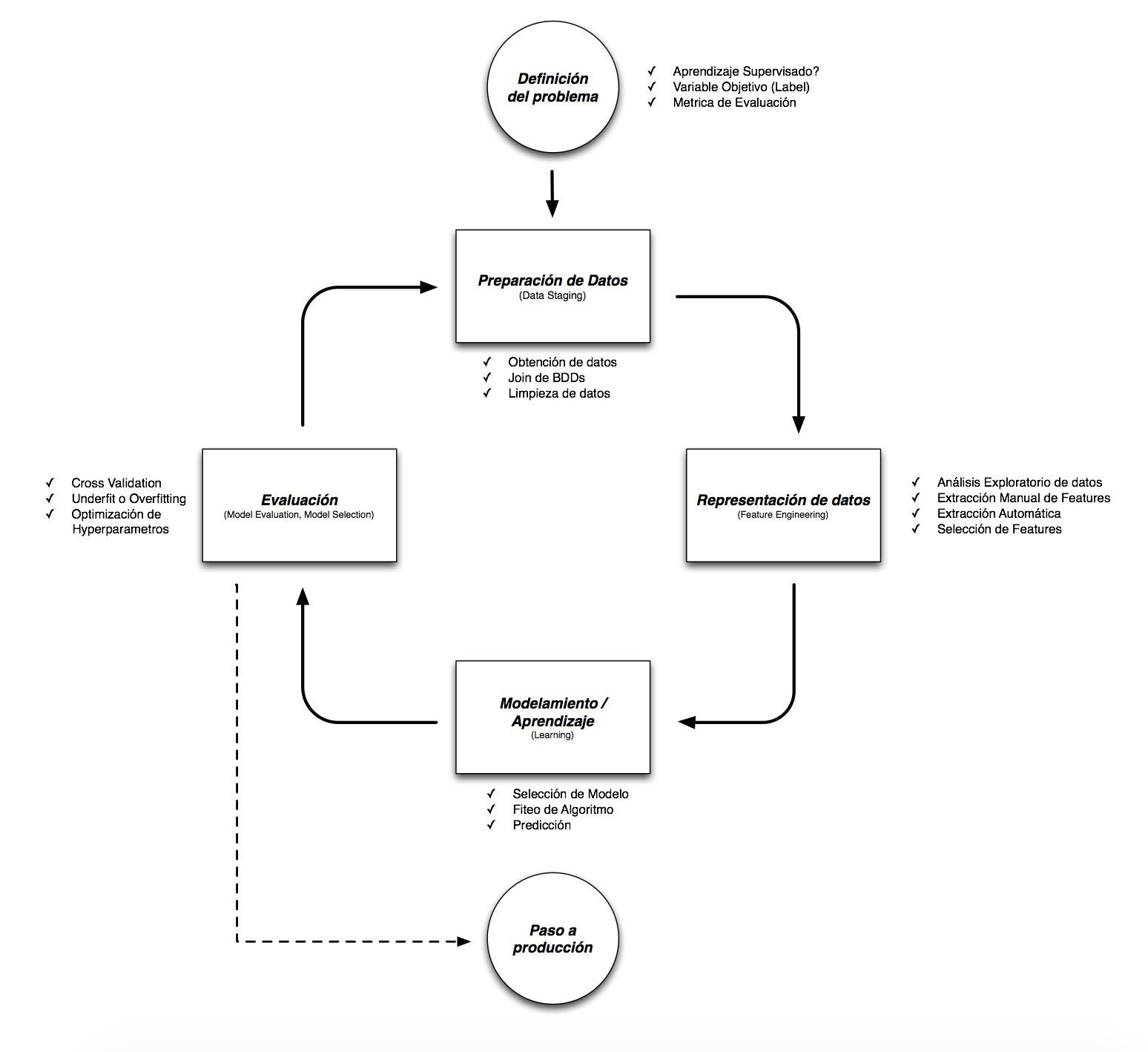
Linea=EjemploL

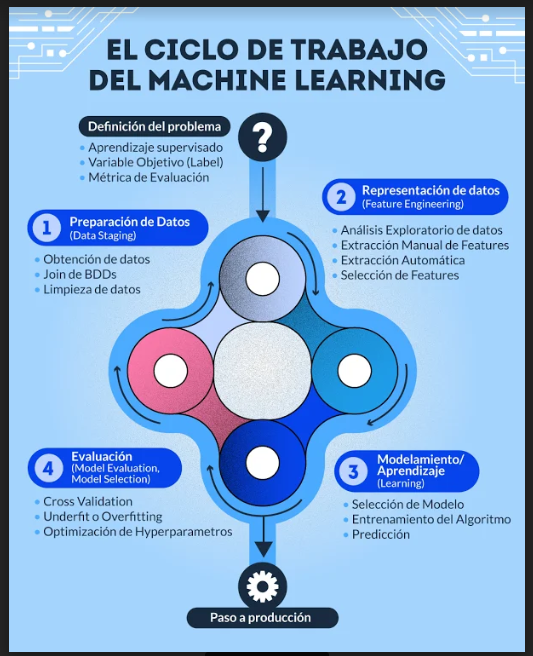
Columna=Feature

Cantidad de columnas=Dimensión de los datos

Output de un algoritmo de ML=Modelo

# Ciclo de vida





# Logaritmos de ML

**Scikit-learn** es una librería libre de Machine Learning para Python.

Presenta varios algoritmos:

* Clasificación
* Regresión
* Agrupación

Incluyendo **SVM, Random Forest, Gradient Boosting, K-Means y DBSCAN**, y es diseñado para interoperar con las librerías **Numpy** y **SciPy**.