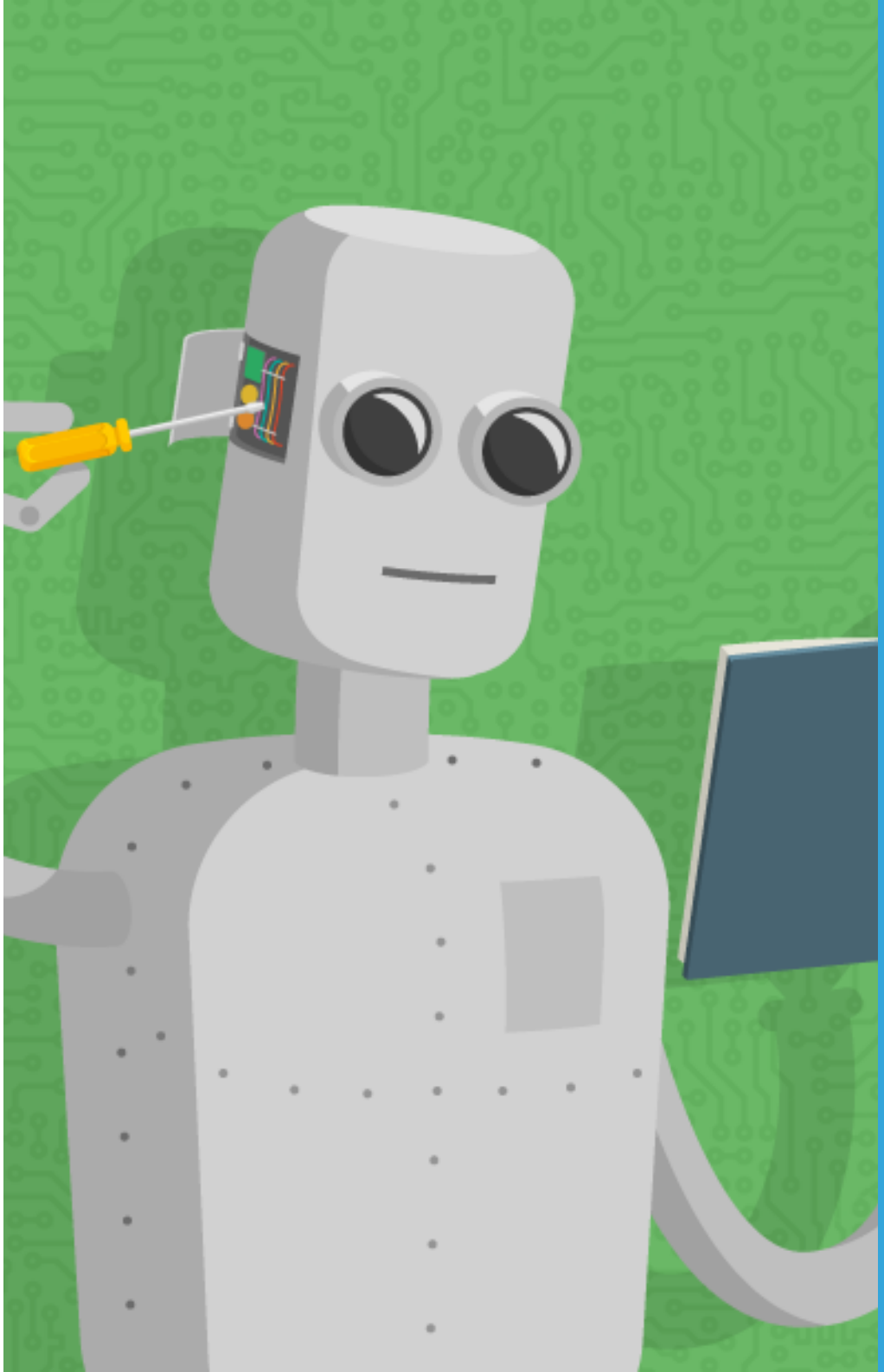
QUORA QUESTION PAIRS COMPETITION

# PROYECTO PNLP



# ADQUIRIR Y APLICAR TÉCNICAS DE ANÁLISIS DE DATOS Y MODELOS PREDICTIVOS MEDIANTE LA FORMACIÓN DE UN PORTAFOLIO DE PROYECTOS KAGGLE

## MISIÓN

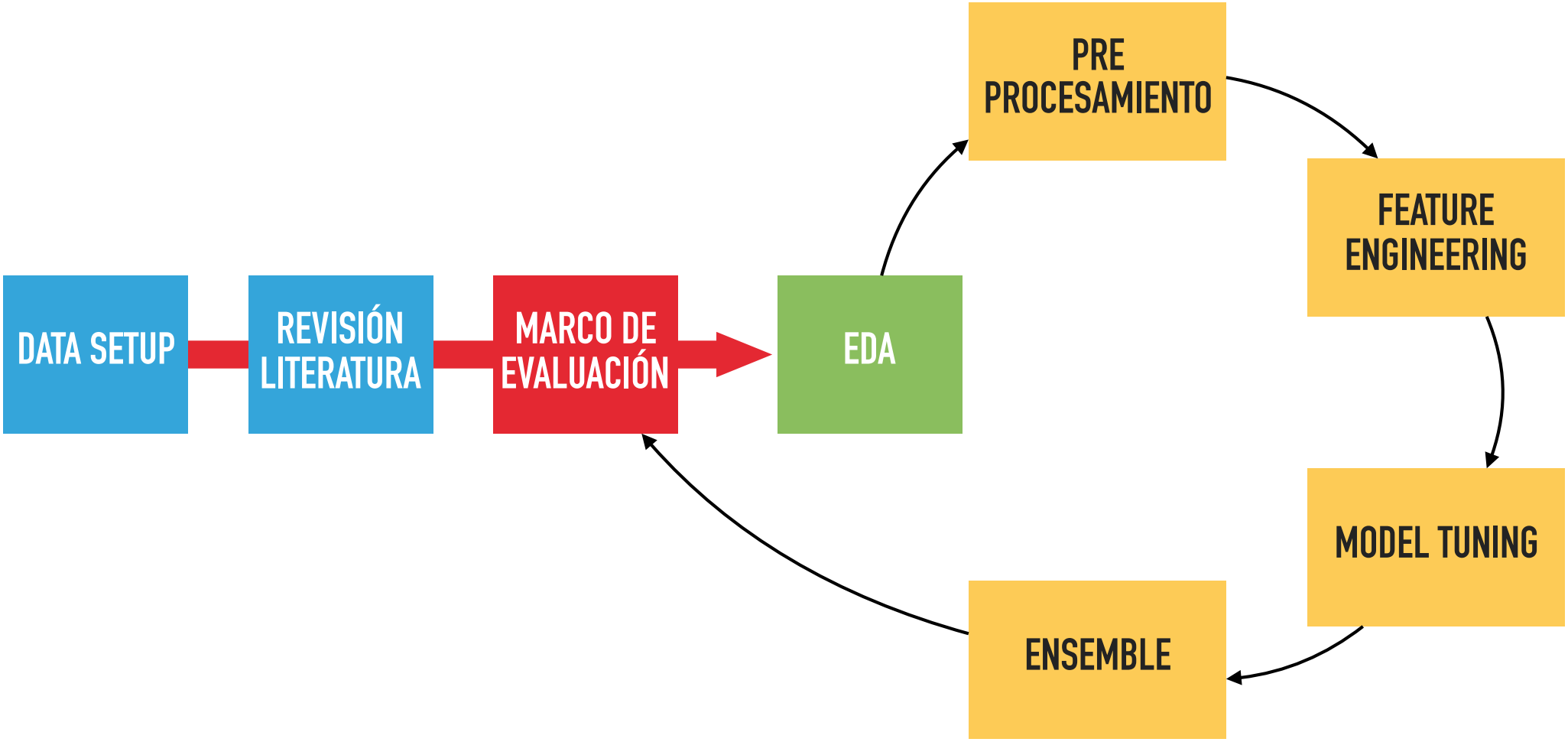
# MISIÓN Y OBJETIVOS

Se espera que al finalizar el proyecto se logren los siguientes objetivos:

- Describir el flujo de trabajo de proyecto Data Science.
- Aprender y mejorar habilidades de programación en R y Python.
- Manejar herramientas estadísticas para la exploración de datos.
- Analizar hipótesis de trabajo e implementar soluciones para corroborar o rechazar estás mismas.
- Comprender la importancia de la definición de métricas de evaluación de modelos, feature engineering, model tunning, y ensembles.
- Conocer y evaluar distintos modelos predictivos relevantes para NLP.
- Implementar procesos de preprocesamientos de datos particularmente para NLP.

# RESUMEN FLUJO PROYECTO DATA SCIENCE

Se compone de varios pasos críticos



# DATA SETUP

El primer paso es disponer una buena calidad de data, esto no es siempre posible, pero en nuestro caso tenemos suerte


**Obtener la data de la competencia "Quora question pairs"**

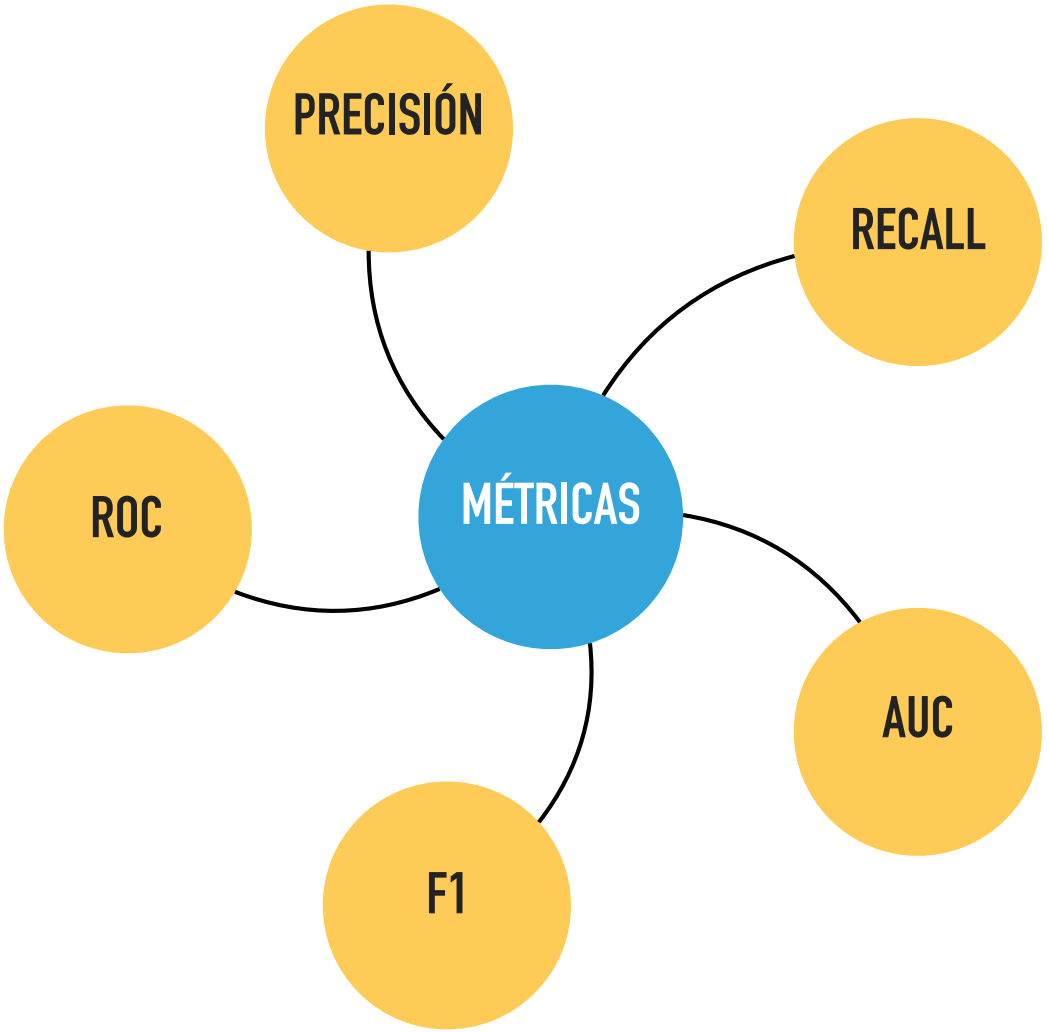
# REVISIÓN DE LA LITERATURA

Generalmente no es necesario reinventar la rueda, alguien mas afuera ya hizo algo semejante (chinito), por lo que es valido gastar tiempo revisando blogs y proyectos mangle anteriores

- www.kaggle.com
- tidytextmining.com
- Data Science workflow
- http://stackoverflow.com
- http://blog.kaggle.com
- Discusión proyecto

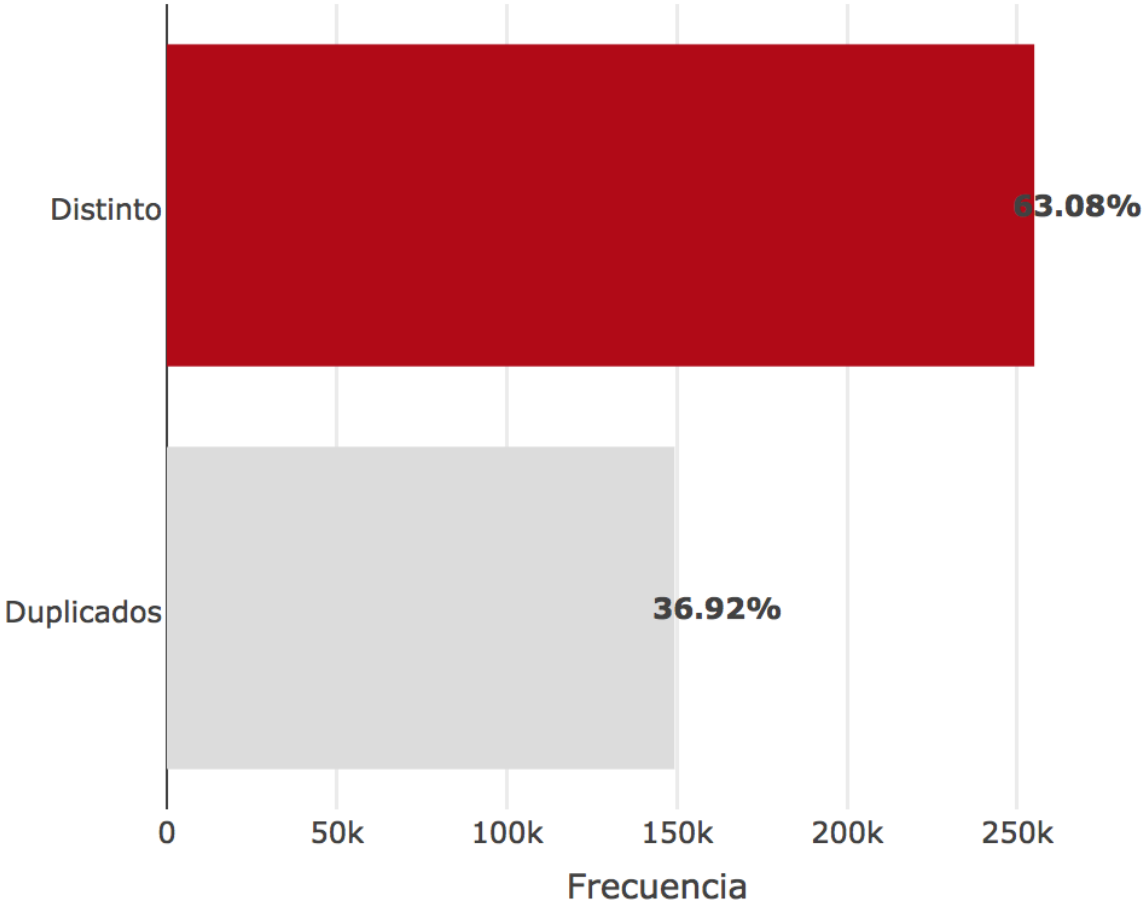
# MARCO DE EVALUACIÓN

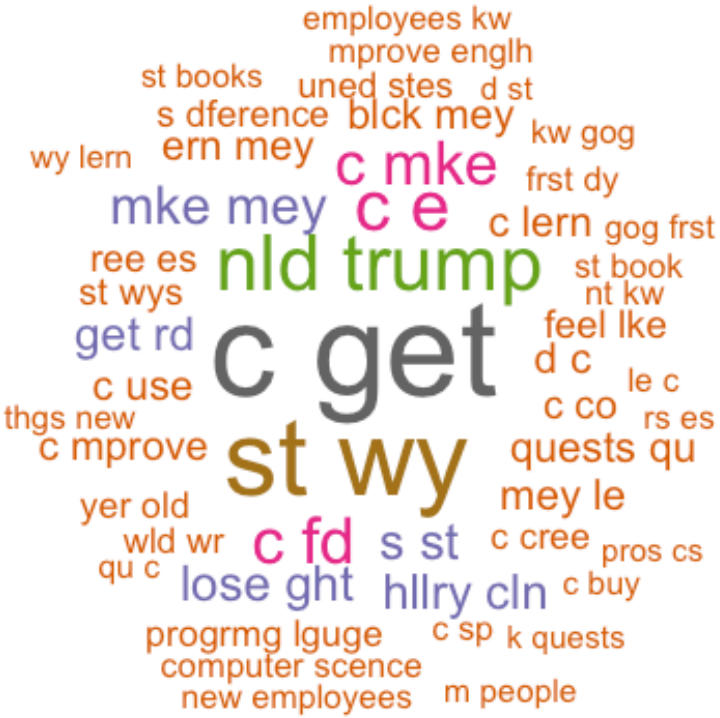
La definición de una adecuada(s) métricas de evaluación es crítico para el desempeño de nuestro proyecto y evitar el overfitting




# EXPLORATORY DATA ANALYSIS (EDA)

Parte fundamental para entender los datos, resumirlos, visualizarlos, y descubrir potencialmente variables que podrían influir en el modelo y procesamiento. Es la base para la construcción de modelos.



# PREPROCESAMIENTO

Es necesario preparar la data para que sea facilmente digerible por el modelo.

- **Fuente**: Integrar de diversas fuentes para mejorar la predictibilidad
- **Calidad**: procesar data ausente, ruidosa o inconsistente
- **Formato**: Datos incompatibles por los modelos, por ejemplo en NLP es necesario transformar palabras en números

# PREPROCESAMIENTO

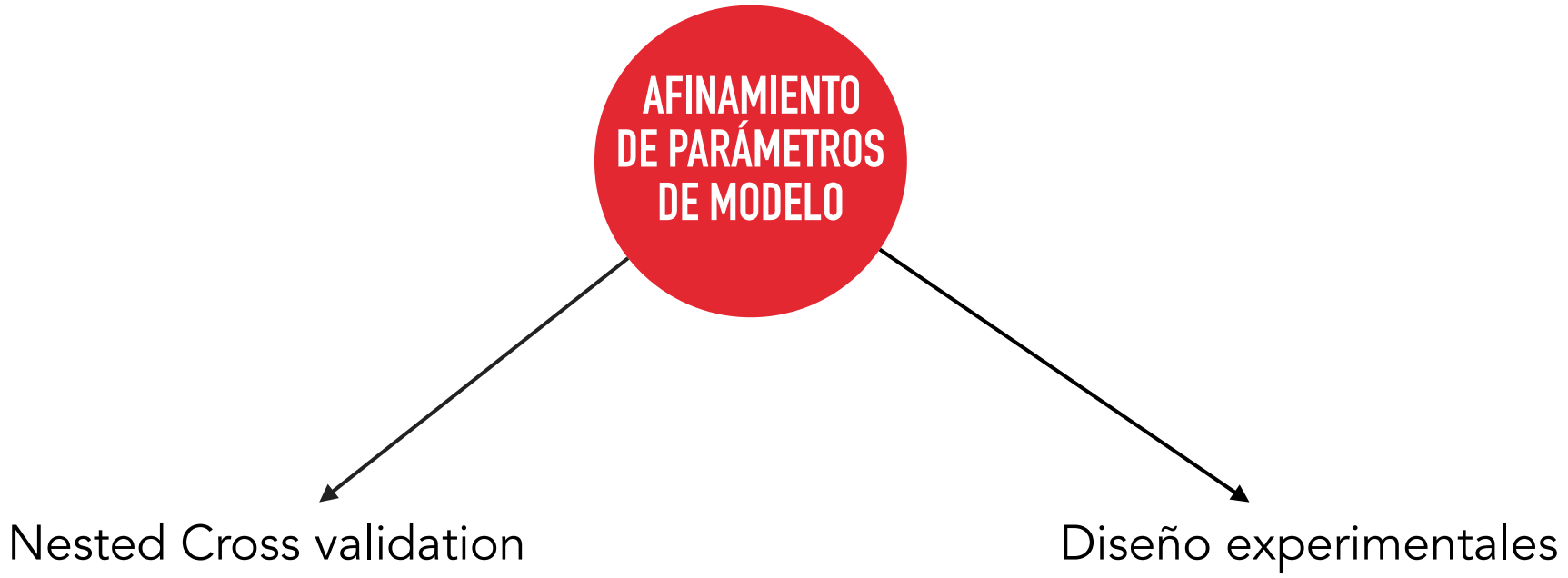
Es necesario preparar la data para que sea facilmente digerible por el modelo.

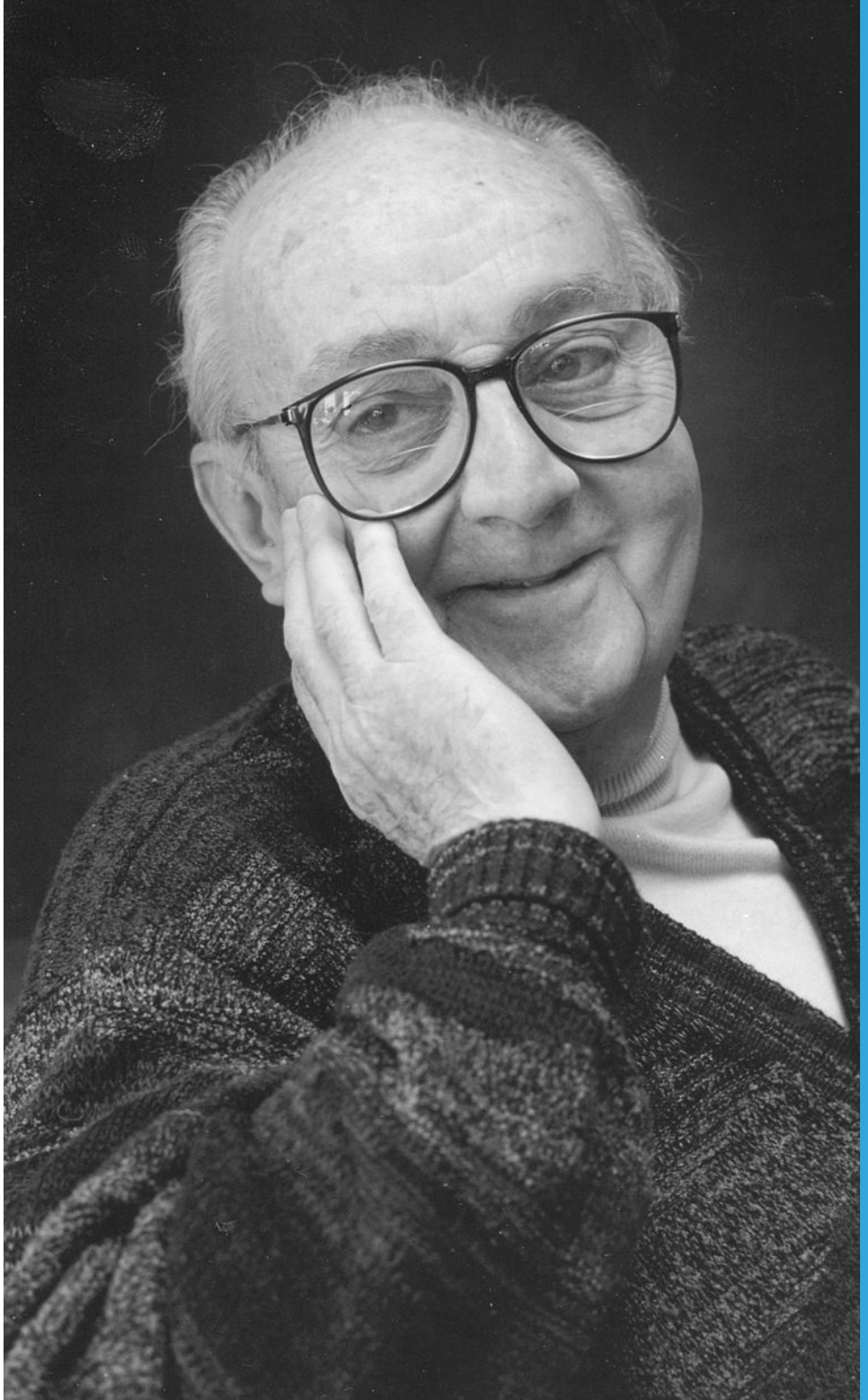
COMING UP WITH FEATURES IS DIFFICULT, TIME-CONSUMING, REQUIRES EXPERT KNOWLEDGE. 'APPLIED MACHINE LEARNING' IS BASICALLY FEATURE ENGINEERING.

FEATURE ENGINEERING

# MODEL TUNING

Este punto es el más simple y es la etapa de afinar los parámetros de los modelos, no obstante puede tomar extensos tiempos de computación





ESSENTIALLY, ALL MODELS ARE WRONG, BUT SOME ARE USEFUL

GEORGE BOX – ENSEMBLE