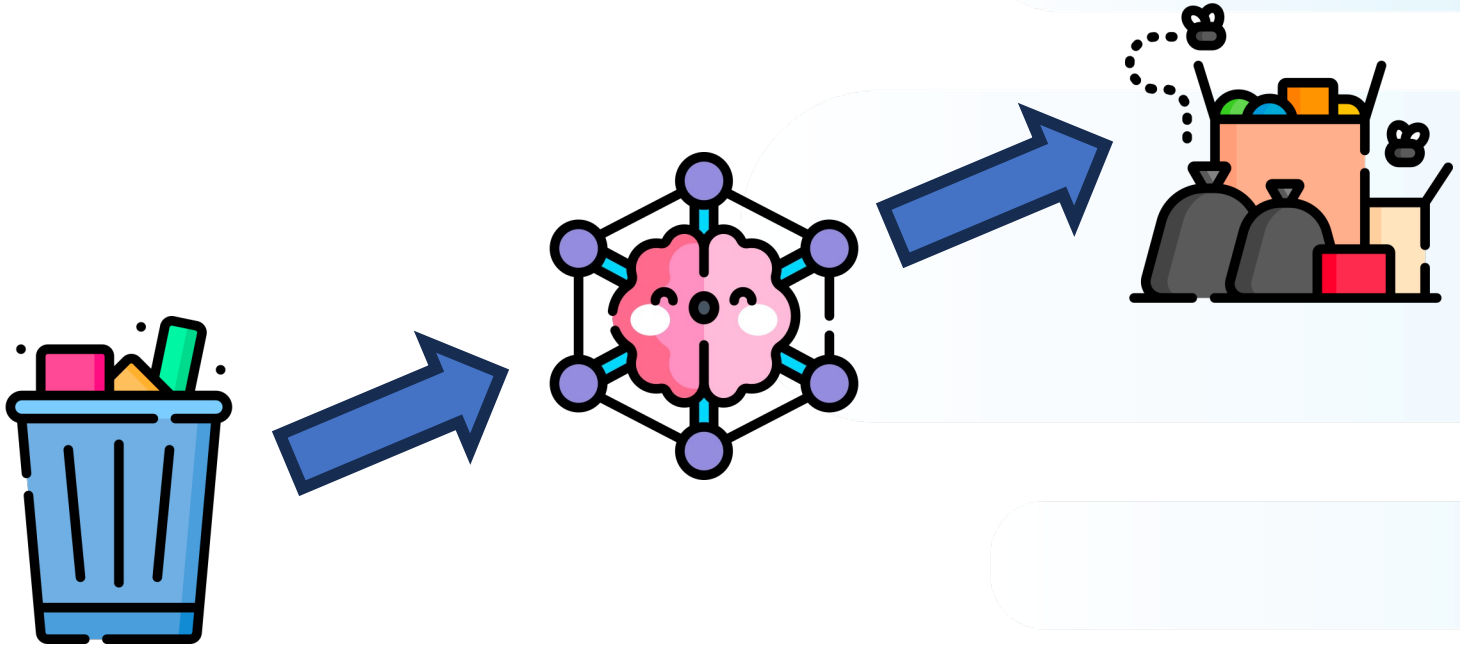
# Validación de datos

# El valor de los datos

- Calidad de los datos ≈ calidad del modelo

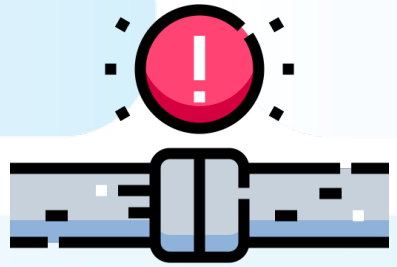
# El valor de los datos

- Calidad de los datos ≈ calidad del modelo



# Validación en un pipeline

- Este paso se encarga de verificar los datos
- Verifica que los datos sea lo que el resto del pipeline está esperando
- En un pipeline cualquier paso tiene la capacidad de detener la ejecución del mismo

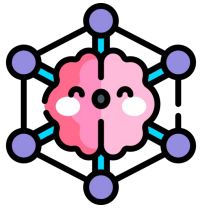


# ¿Qué hay que validar?

- No anomalías en datos
- El schema (o esquema)
- Coincidencia con las estadísticas de los datos anteriores

# ¿Cómo se lleva a cabo?

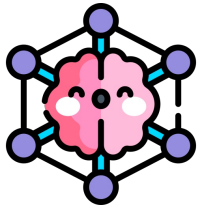
Desarrollo



Modelo

# ¿Cómo se lleva a cabo?

Desarrollo

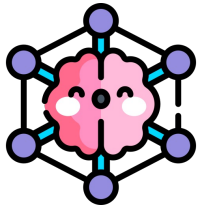


Modelo

Features

# ¿Cómo se lleva a cabo?

Desarrollo



Modelo

Estadísticas

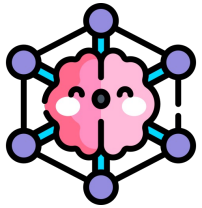
Features

Schema

Pipeline en producción

# ¿Cómo se lleva a cabo?

## Desarrollo



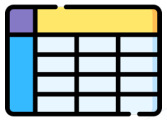
Modelo

Estadísticas

Features

Schema

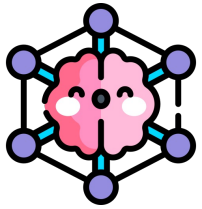
## Pipeline en producción



Nuevas features

# ¿Cómo se lleva a cabo?

## Desarrollo



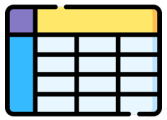
Modelo

Estadísticas

Features

Schema

## Pipeline en producción



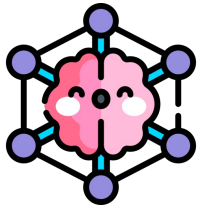
Nuevas features

Estadísticas

Schema

# ¿Cómo se lleva a cabo?

## Desarrollo



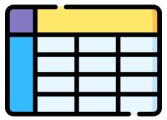
Modelo

Estadísticas

Features

Schema

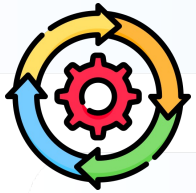
## Pipeline en producción



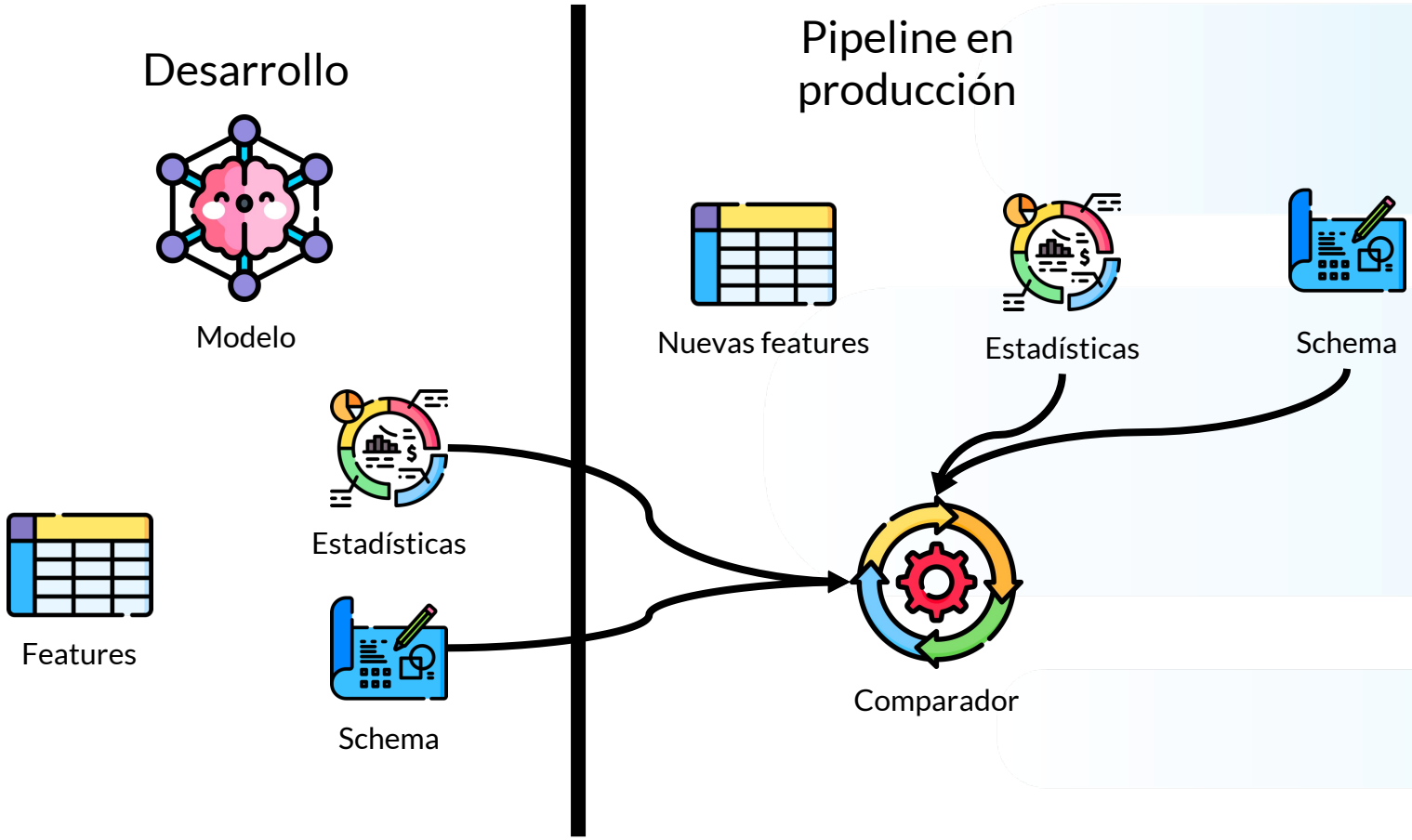
Nuevas features

Estadísticas

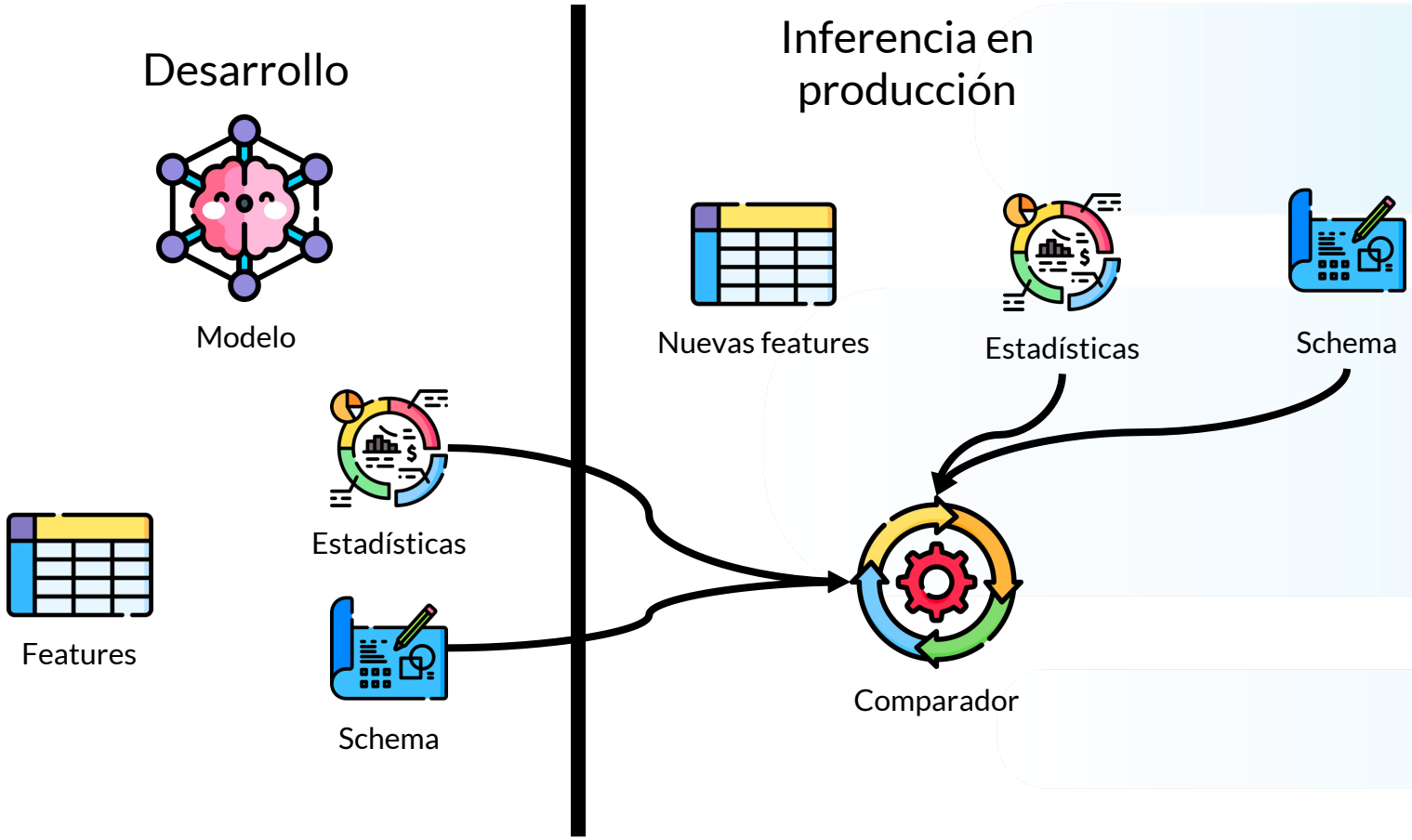
Schema



# ¿Cómo se lleva a cabo?



# ¿Cómo se lleva a cabo?



# Consideraciones

- Distribuye la validación de datos si el dataset es muy grande
- Considera validar solamente un subconjunto si el dataset es muy grande
- A la validación de los datos puede ser usada para monitorear