Pipeline de productos de datos

Procesos

- Regularmente existen varios pasos de procesamiento para preparar los datos.
  - Extraer los datos (desde una carpeta, el internet, una base de datos) e importarlos al hdfs.
  - Validar los datos.
  - Transformarlos a un formato más adecuado.
  - Ejecutar agregaciones y generación de variables.
- Y pasos para preparar el modelo
  - Entrenar, validar y seleccionar modelos.
  - Poner en producción el modelo seleccionado

Procesos

- Además estos pasos se empiezan a ejecutar cuando:
  - A un tiempo dado
    - e.g. Cada medianoche
  - Un evento ocurre
    - e.g. Se agregó un nuevo archivo
- Coordinar los pasos
  - Un paso se sigue al otro, sólo si el anterior terminó exitosamente.
  - Repetir el paso
- Tomar acciones de gestión
  - Mandar correos
  - Tomar tiempos de ejecución

Orquestación

- Al concepto de coordinación, gestión, programación se les conoce como **orquestación**.
- La orquestación (como muchas cosas en ciencia de datos) se representa por un grafo dirigido acíclico (**DAG**).
  - Los `RDDs` de `Spark` también son un DAG.
- A un **DAG** se le conoce como *workflow* y a la administración de los *worflows* se le conoce como orquestación de *workflows* (también conocidos como *pipelines*).
- En esta clase veremos a **`Luigi`** como orquestador.

¿Qué es un pipeline?

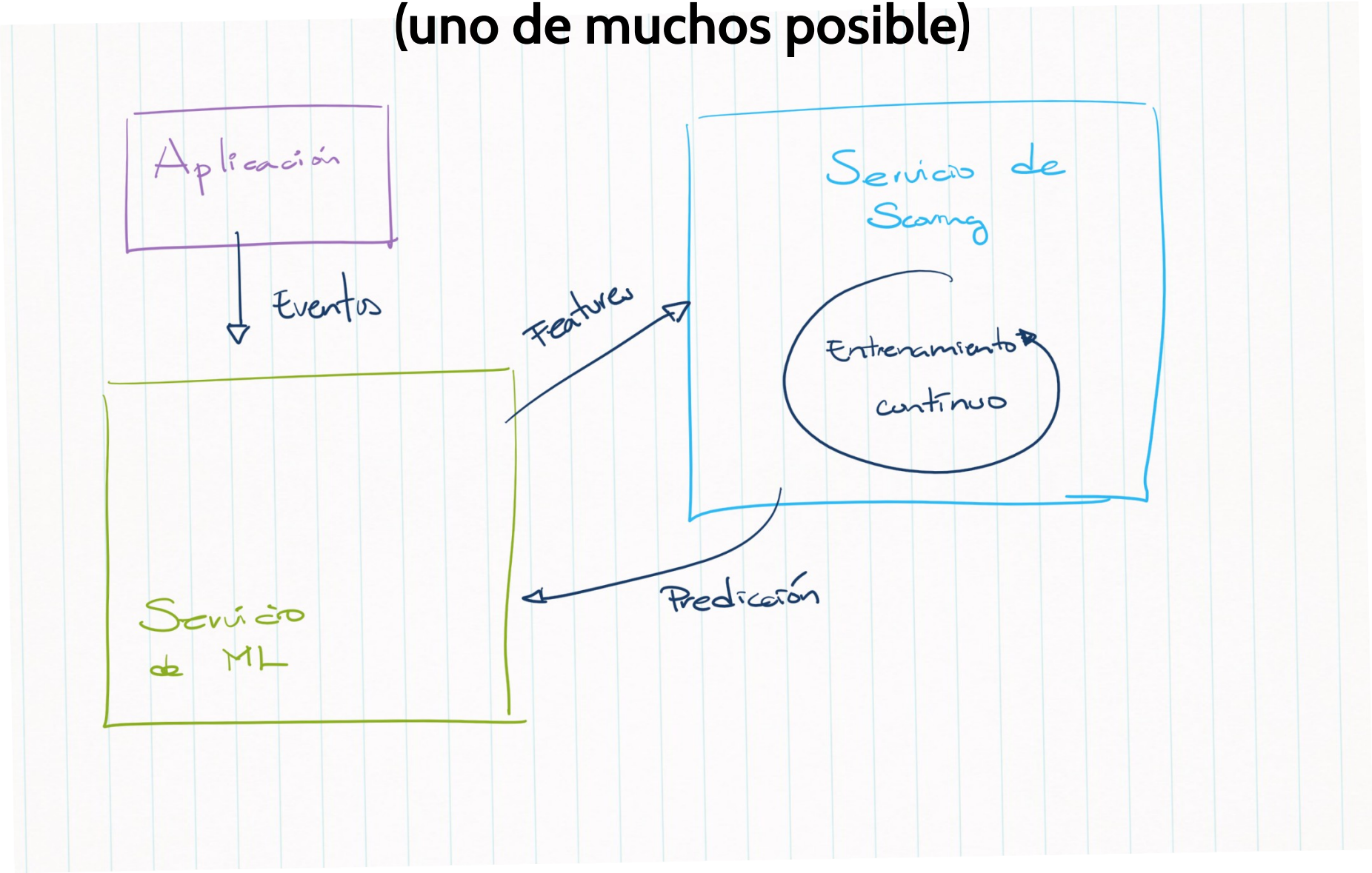
- Se aplican una serie de transformaciones a los datos
- Los *pipelines* definen una línea de herencia de los datos
  - *Data lineage*
  - siempre tenemos los datos crudos para reejecutar.
- Es como armar legos
  - Flexibilidad, Modularidad, *Testability*

¿Luigi?

- Orquestador de Spotify
- Escrito en python.
  - Cualquier cosa funcionará entonces: scikit, pyspark, etc.
- Integrado con Hdfs.
- Soporta *out-of-the-box* postgresql
- Soporta **Idempotencia**
- Soporta *checkpointing*

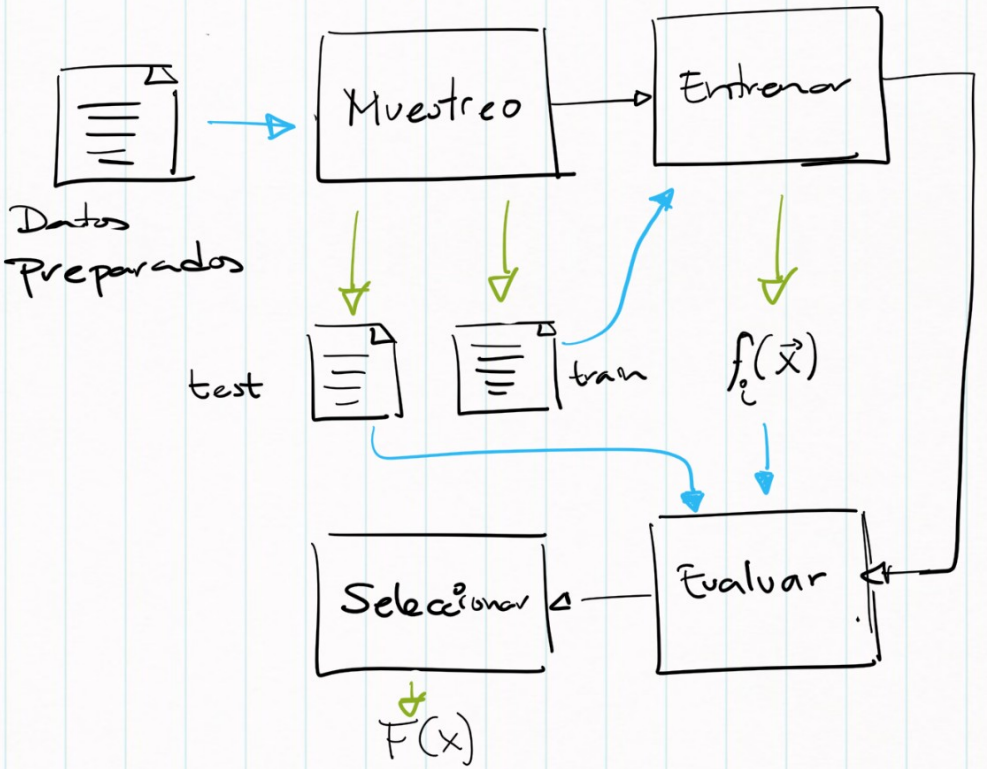
¿Cómo se ve un producto de datos?

(uno de muchos posible)



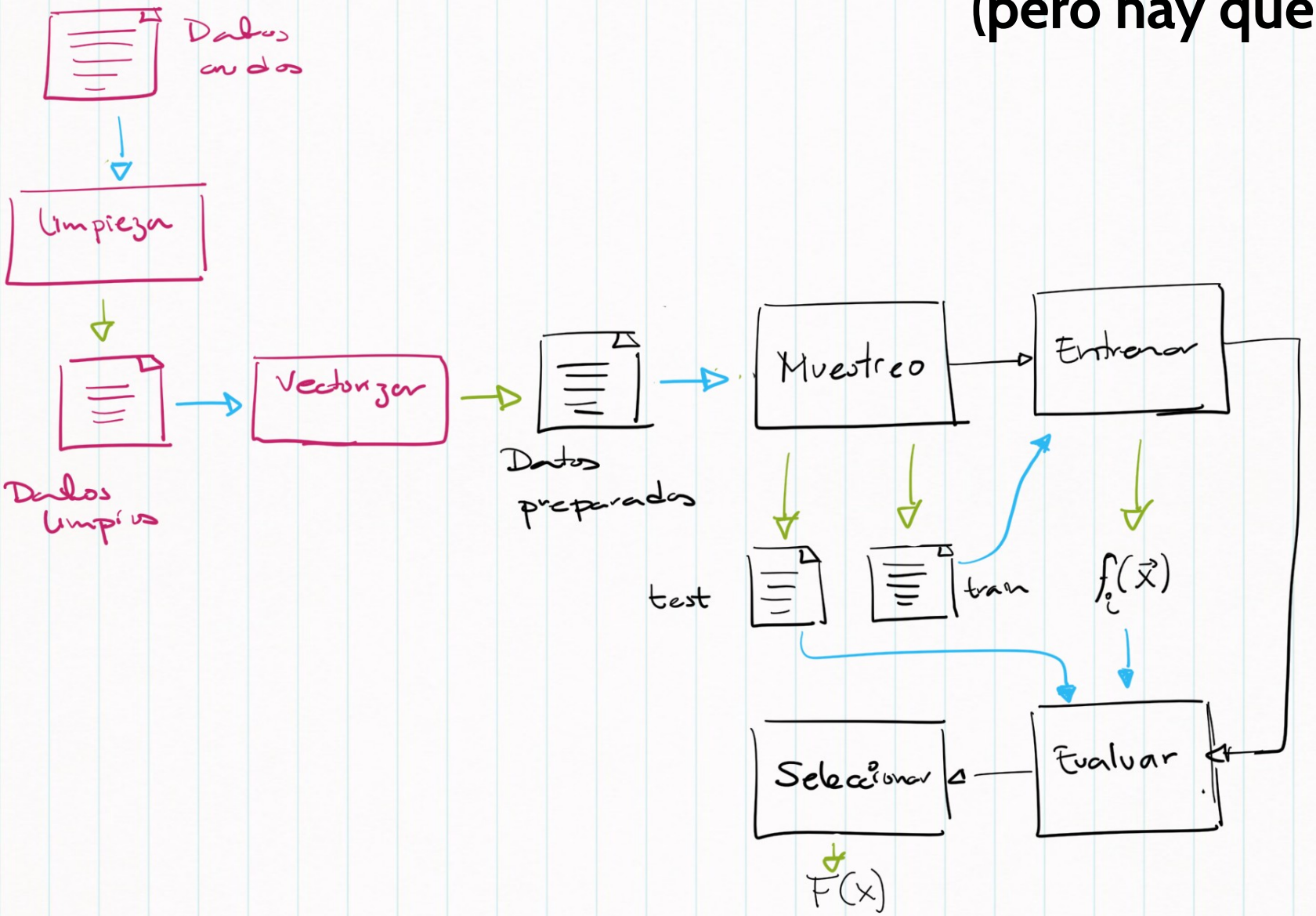
El proceso de modelar

(regularmente se hace a mano)

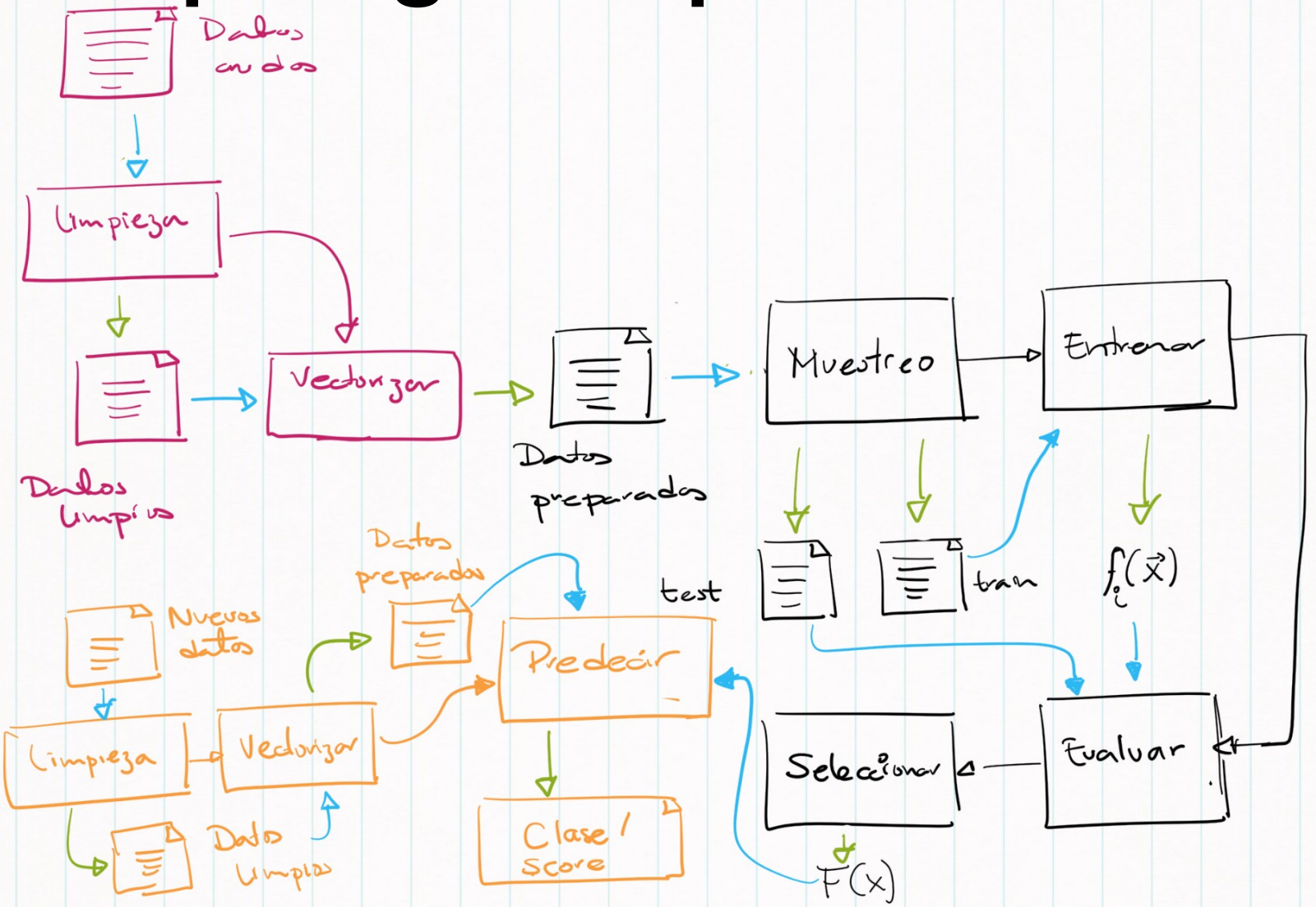


Lo que no queremos hacer...

(pero hay que hacer)



Lo que significa producto de datos



Demostración