Tres ideas en torno a la reproducibilidad de los procesamientos geodésicos.

Javier José Clavijo

jclavijo@fi.uba.ar
https://github.com/jjclavijo/3ideas2022sirgas

Universidad de Buenos Aires - Facultad de Ingeniería

La charla habla principalmente sobre ideas para mejorar los flujos de datos para que sean más reproducibles (y varios conceptos asociados que ya veremos).

El hilo conductor es la situación que me motivó a generar este trabajo.

Habiéndome propuesto aplicar métodos bayesianos al calculo del marco de referencia a partir de soluciones diarias o semanales caí en cuenta que lo primero que tenía que entender bien era cuales son los supuestos del modelo y que distribuciones a-priori implican, incluyendo las correlaciones entre todos los parámetros que entran en juego.

Aquí topamos con algunas dificultades a la hora de analizar y comparar los procesamientos publicados. Se hace dificil decidir si determinado conjunto de datos publicado es apto para un objetivo concreto, porque no podemos saber si hay correlaciones no declaradas que nos puedan afectar el resultado si no sabemos cómo se relacionan con sus datos de entrada. (o no sabemos qué datos de entrada tiene)

Puesto más formal, necesitamos saber si las distibuciones de los parámetros provienen de condicionar unos a otros, marginalizar o si en algun caso ambas cosas son equivalentes (es decir que los parámetros son independientes).

Ya vamos a ver por que importa marginalizar o condicionar.

La principal dificultad para identificar esto es la gran variedad de software y las dificultades para seguir exactamente el proceso que hacen.

Incluso el proceso de lectura de datos y la forma en que los manejan es diferente en todos. Por ejemplo, las estructuras de lectura y uso de datos de RTKLib y G-Nut son radicalmente diferentes. (no solo por las diferencias en la lógica de C y C++)

Esto aveces dificulta que se pueda aprender los procesos asociándolos a lo que está pasando por detrás en lugar de generar “rutinas” y “automatizarlo”.

Como consecuencia de este problema surge la idea de plantear esquemas par describir los procesos y flujos de trabajo.

Aquí se abren dos ramas importantes.

Por un lado los estándares de “data provenance” que indican formas de describir los procesos en forma sistemática.

Como introducción ver los links que siguen:

- https://www.w3.org/TR/prov-dm/
- http://videolectures.net/iswc2014_gil_semantic_challenges/ minuto 18:00 en adelante

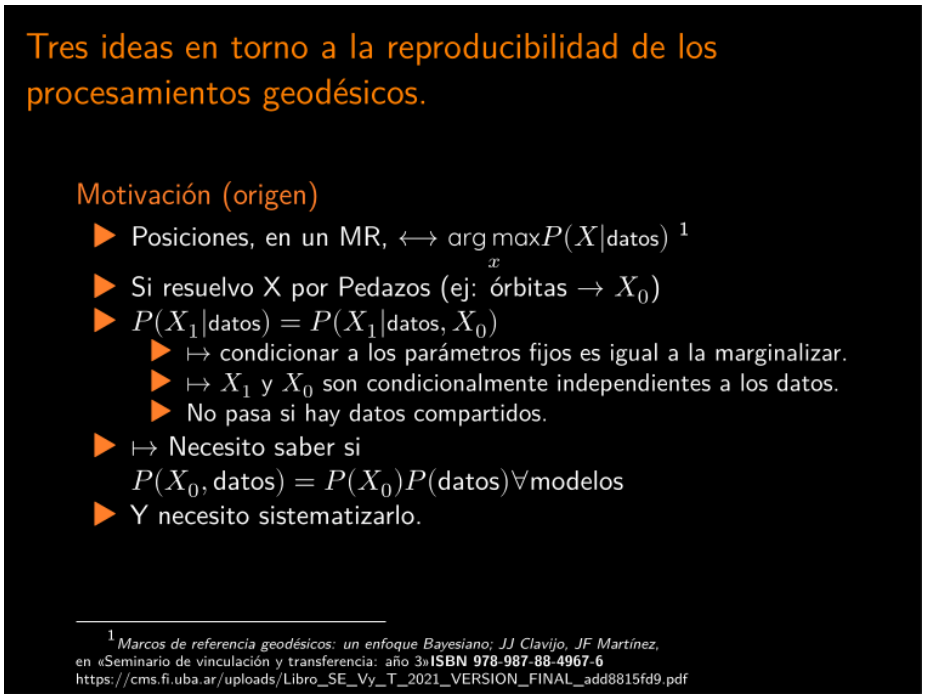

Figure 1: Pagina 1

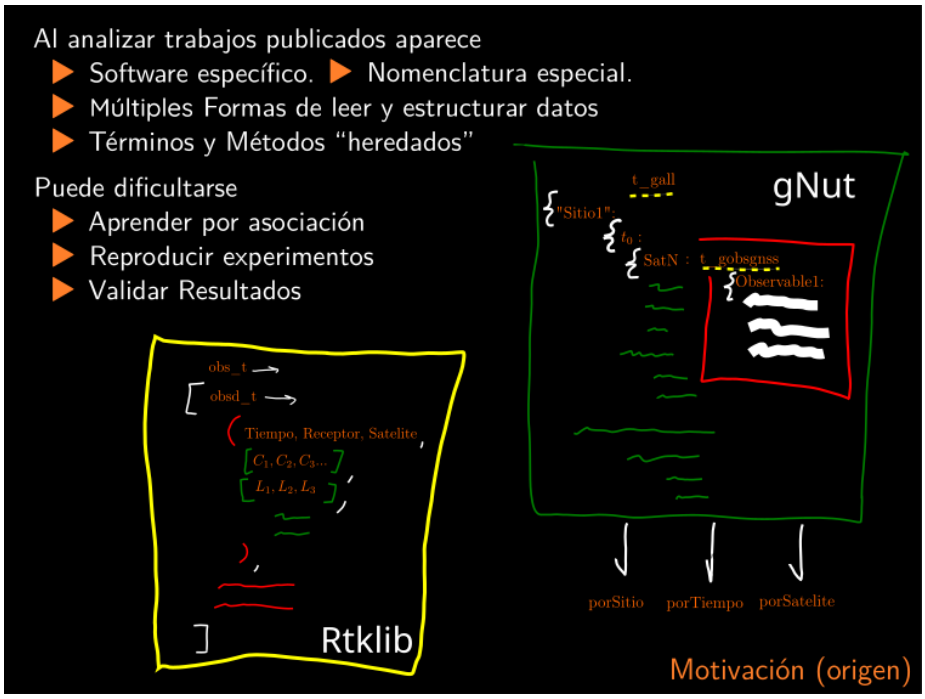

Figure 2: Pagina 2

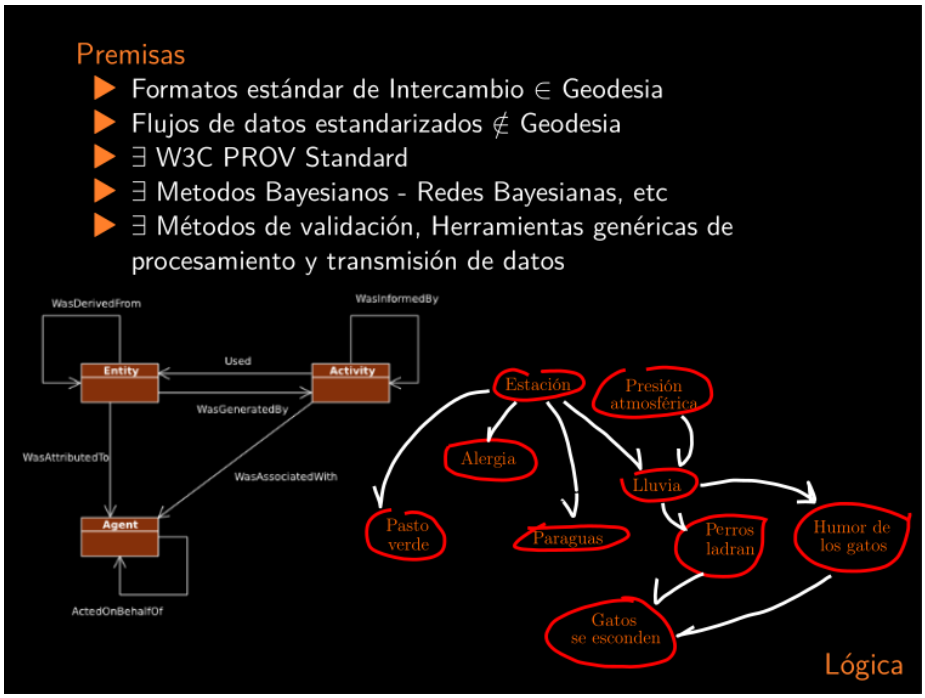

Figure 3: Pagina 3

Por otro lado, las redes bayesianas son una forma de describir procesos modelos estadísticos en función de relaciones causales. Podemos ver que, si conjuntos de datos comparten relaciones causales con algún juego de parámetros no deberíamos tratarlos como si fueran independientes sino sólo condicionalmente independientes.

En cuanto al procesamiento GNSS, que tomamos como ejemplo, podemos hacer algunas observaciones adicionales.

1. Los datos de observación pueden representarse como un tensor ralo. Cada modelo paramétrico que contribuye al modelo de observación parte de un espacio de parámetros bien acotado y genera un tensor denso de la misma dimensión que el tensor de observaciones.
2. Los resultados son generalmente estimadores de los parámetros que se consideran como si provinieran de una normal multivariada.

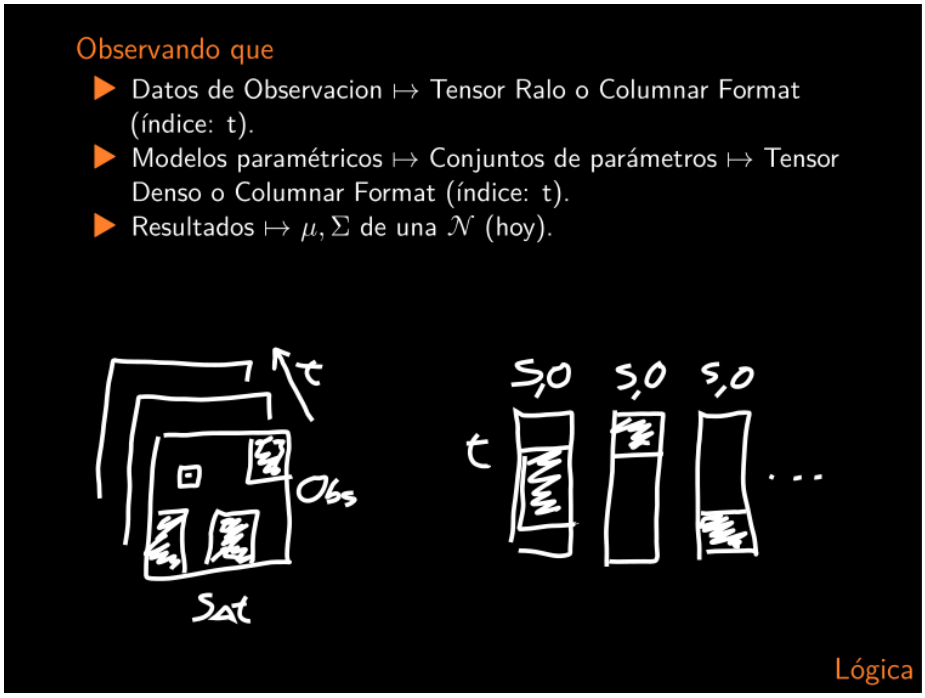

Figure 4: Pagina 4

Esto nos da dos líneas generales importantes para describir los procesos en función de las transformaciones que sufren los tensores y las formas en que los parámetros se relacionan con los tensores que generan.

La propuesta es entonces, pensando las entradas como segmentables en el tiempo, (lo cual adicionalmente permitiría un esquema de paralelización), describir los flujos de trabajo en forma funcional (misma entrada = misma salida), y usar las herramientas existentes para generar formas de validar los datos y los productos generados, identificar y describir las dependencias, etc.

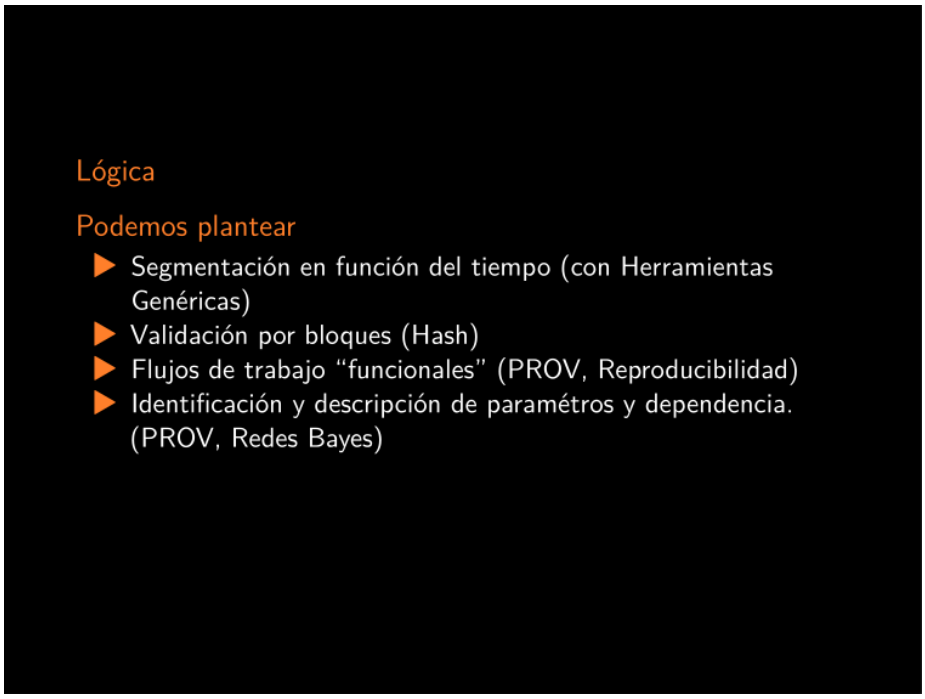

Figure 5: Pagina 5

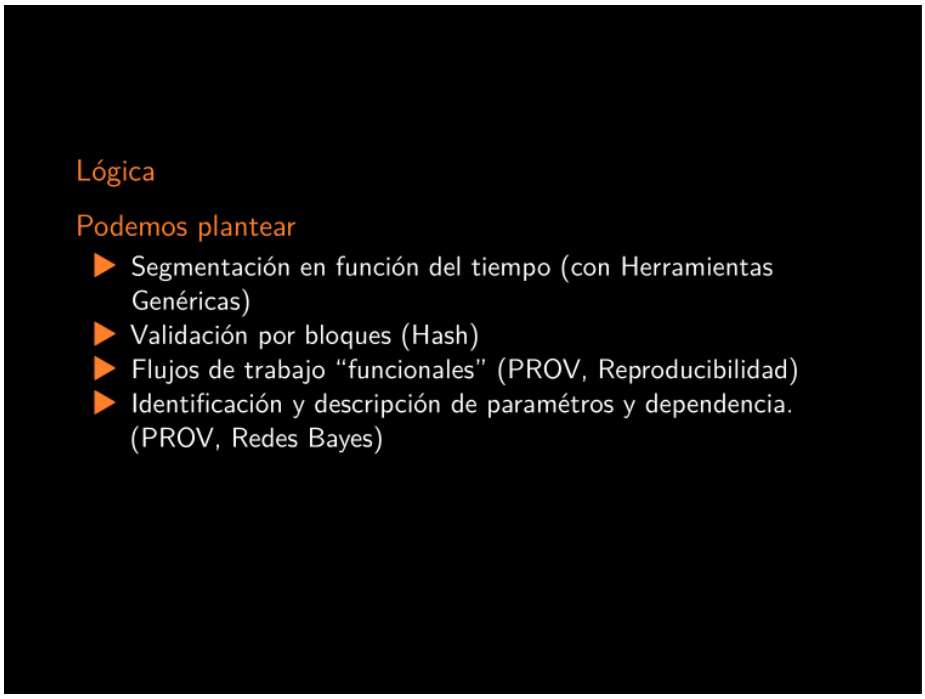

Figure 6: Pagina 5

Tenemos en geodesia multiplicidad de observaciones, muchos procesos y muchos modelos.

En general, hay algunos modelos cuyos parámetros damos por dados y otros sobre los que aplicamos algún método de inferencia.

Lo ideal sería que los parámetros no tengan relación entre si (arriba), pero termina resultando que pueden tener relación a través de los datos que usamos para hacer inferencia (abajo).

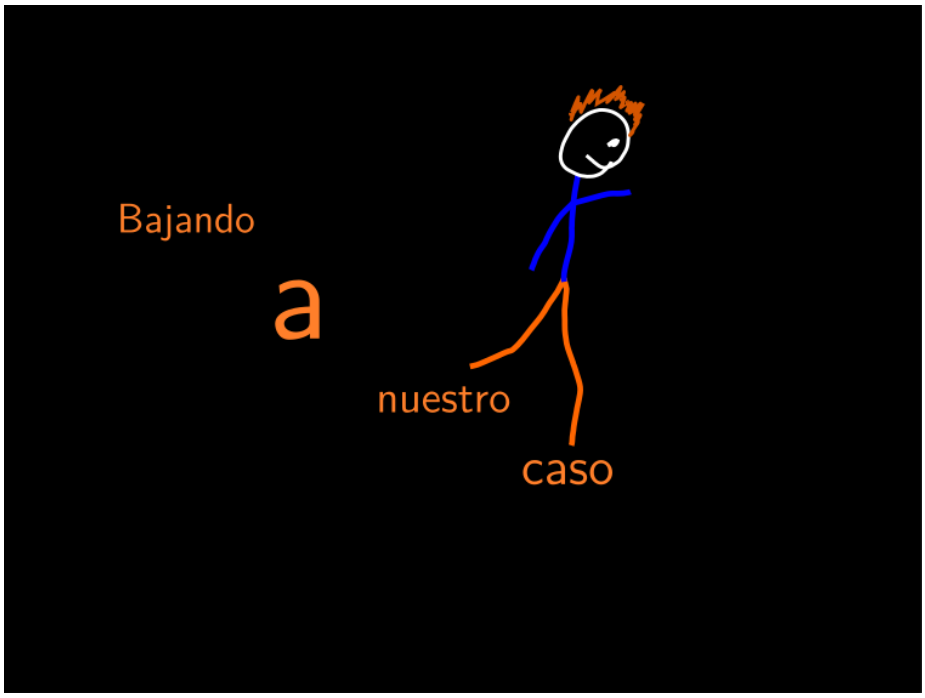

Figure 7: Pagina 6

Y el problema de la inferencia de parámetros no termina ahí, porque muchas veces los métodos de inferencia utilizados no son exactamente idénticos entre diversas implementaciones de software, o tienen hiperparámetros que no están debidamente documentados, etc.

Por ejemplo, cuando estamos usando un filtro Kallman, mas allá de las multiples variantes de este, ¿qué significa "fijar" ambiguedades? ¿Se las ignora en el filtro o se fuerza el resultado?

Los problemas siguen (cerrando el planteo del caso de ejemplo), porque la distribución de los datos es poco homogénea en general, lo que hace que sea difícil producir una validación real. Hay estaciones que terminan influyendo en todas las soluciones, aunque sea a través de la estimación de algunos de los parámetros que se fijan al procesar.

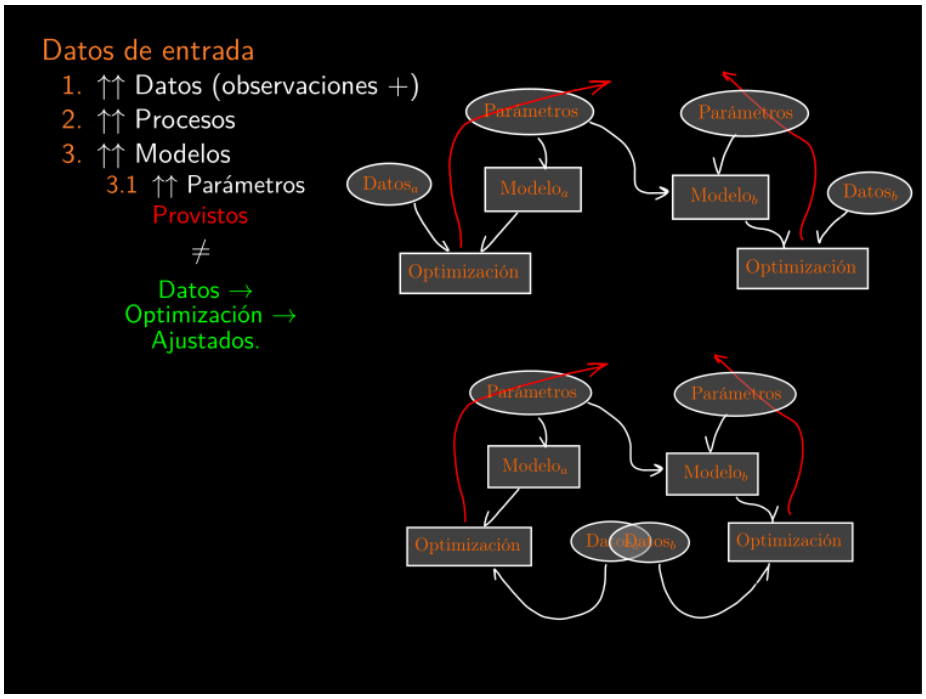

Figure 8: Pagina 7

Distribución de los datos

1. No es homogénea
2. Difícilmente permite validar con conjuntos independientes
3. Puede haber correlaciones ocultas a través de parámetros “dados por hecho”.

Tres ideas introduciendo

los conceptos de

1. Funcionalidad
2. Replicabilidad.
3. Reproducibilidad.
4. Robustez.
5. Validación.
6. Rastreo de origen. (data provenance)

Idea 1

Aplicar el estándar PORV-O

Agente - Entidad - Acción

- Esquema funcional (= entradas $\mapsto$ = salidas)
- Red Bayesiana

Aplicar el estándar PORV-O

Ejemplo: Analizar algoritmos de detección de discontinuidad

Leer 150 papers para saber

- ¿En qué MR trabajan? (como hiperparámetro al menos)
- ¿Qué tipo de modelo aplica? (Lineal?)
- ¿Hasta qué distancia “detectan”?

Ej de provecho: Modelos lineales no dependen de cambios lineales del MR

Caso 1:

Coordenadas Nevada -> Modelo lineal

Caso 2:

Coordenadas Sirgas -> “realineación” -> Modelo paramétrico

Caso 3:

Procesamiento -> Alineado a Sirgas -> Modelo…

Caso 4:

Mi Caso

Idea 2

Planteo

El formato RINEX es texto ↦

- No define tipo de dato estandarizado para cada variable.
- Depende de un parser que puede ser mas o menos tolerante.
- Espacios y ceros pueden “cambiar” el archivo sin cambiar el dato.

Datos transmitidos por NTRIP También son (otro) texto.

BINEX

- Mejora todo esto, pero no es Columnar y los tipos de dato no son estándares de computadora (sigue habiendo algo de nicho)…

Planteo

La naturaleza del dato es Observaciones de una variedad de indicadores indexados por el tiempo

- Se puede estandarizar en columnas con tipo de dato determinado

---

Utilidades

- Al fijar el tipo de dato se puede comparar datos (no archivos) con hashes.
  - (BINEX podría permitirlo, aunque mezcla muchos indicadores “extra”)
- Se podría partir y transmitir los datos eficientemente. (aplicaciones RT) (y validarlos)
- Se puede usar entornos como spark, arrow-flight, etc.

Hashes

https://github.com/jjclavijo/sirgas2022hashes

Columnar Format: Flujos de trabajo

https://github.com/jjclavijo/sirgas2022arrow

Idea 3

Planteo

Es muy difícil garantizar la independencia real de procesamientos, aún si separáramos redes que no se conectan.

Idea 3

Observación

Los productos publicados no describen su propio origen (condicional o marginal).

Ejemplo: Los SINEX pueden ser subsets de otros SINEX: condicionales $X_1 = X|X_0$ o marginales $X_1 \ni P(X_1) \times P(X_0) \neq 0$

Los sp3 dan parámetros (θ) de órbitas pero no describen $P(\theta, x_0)$

Si nuestros parámetros (X_2) son tales que $X_0 \cap X_2 \neq \emptyset$ estamos perdiendo información, porque no sabemos nada de P(X_0).

Observación

- El SINEX de SIRGAS no describe su dependencia del SINEX de óbitas de IGS, de hecho lo ignora porque usa los sp3 para calcular.
- Rastreando el origen de datos podría rastrearse esa dependencia (desde el SINEX de IGS)
- Esto permitiría diseñar la combinación de forma distinta usando una red bayesiana y Beleaf propagation

---

Ejemplo

Dependencias SINEX bayes net. SIRGAS CODE IGS

Validación total (?)

Una idea intereseante:

- Tener conjuntos de órbitas independientes.

- Validar procesamientos con verdadera independencia modelos / datos.

¡Muchas Gracias!